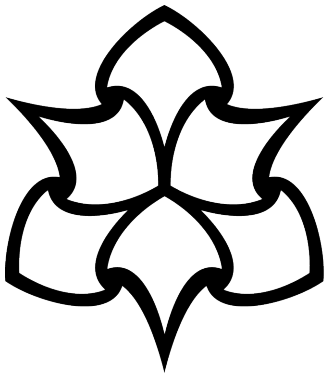


Developments in Next-Generation Sequencing and Bioinformatics for Ecological Genetics

G FOX

PhD 2019

Developments in Next-Generation Sequencing and Bioinformatics for Ecological Genetics

Graeme Fox

A thesis submitted in partial fulfillment of the requirements of

Manchester Metropolitan University for the degree of Doctor of Philosophy.

Ecology and Environment Research Centre,

Department of Natural Sciences,

Manchester Metropolitan University.

2019

Contents

Abstract

This thesis investigates the applications of next-generation sequencing to ecological studies by interrogating the power of high-throughput sequence data and developing resources to better understand the capabilities and limitations.

In Chapter two, I use microsatellite detection methods to develop new markers for *Raja undulata* and show that after the production of the first captive generation, this small population shows no evidence of inbreeding depression or effects of genetic clustering by aquarium. I demonstrate the population has retained high genetic diversity throughout and highlight the importance of genetic management of *ex situ* populations.

In Chapter three, I develop a novel *in silico* microsatellite marker design method. This new method allows the automated removal of markers likely to show elevated rates of null alleles, allelic dropout or cryptic fragment length altering mutations which invalidate the assumptions of mutation at a microsatellite locus. Furthermore, the method enables the automatic selection of likely polymorphic loci, thus removing many of the inefficiencies of marker design.

In Chapter four, I perform parallel microsatellite and single nucleotide polymorphism (SNP) analysis to compare the application and relative power of each marker type in the analysis of the population structure of the larval dispersing decapod, (*Homarus gammarus*). Neither marker detects any genetic structuring in the fisheries of the UK and Ireland implying that genetic mixing is extremely high. SNP analysis is the preferred method due to quicker generation of data and results.

In Chapter five, I conduct an investigation into the biases involved when selecting a metabarcoding marker for analysis of plant communities in mixed pollen samples collected from honey bee hives. I find high rates of false-positive identifications and

highly contrasting descriptions of plant communities, indicating low confidence in the data generated by each individual marker. I conclude that for plant metabarcoding, multiple parallel markers are required to improve confidence in individual taxa calls, and to broaden the detection range; important where highly cultivated gardens are accessed as well as the native flora.

Finally, I conclude the thesis with a general discussion of the methods and findings of the previous chapters and discuss the merits and drawbacks of the methods employed.

List of Tables

List of Figures

Declaration

I hereby declare that the work has been done by myself and no portion of the work contained in this thesis has been submitted in support of any application for any other degree or qualification on this or any other university or institution of learning.

Acknowledgements

A great many people have helped me to get to the point where I am able to write this thesis. My greatest thanks go to Jenny Rowntree and Richard Preziosi who have offered countless hours of their time and expertise to guide me through my PhD over the last five and a half years. As my supervisors, I have had the great benefit of their support, patience, and incredible scientific knowledge, and will always be grateful that I was able to undertake this PhD in their group(s). This PhD has spanned two universities and I must thank all my colleagues at both The University of Manchester (UoM) and at Manchester Metropolitan University (MMU) for their advice, great humour, and all-round support to keep going. In particular, Paul Fullwood, Chantal Hillarby and Caroline Grimshaw at UoM were all instrumental in encouraging me to pursue a PhD, and very literally made it possible. Liam Campbell (and Steve Canty when in the country) could *always* be relied upon for a beer when stress made it necessary. I must thank all the members of the Rowntree, Preziosi and Harris labs for their great scientific insight, statistical knowhow, lab wizardry and willingness to share all of the above. In particular Sarah Griffiths has been a font of advice in most aspects of this work. My friends and family have been an incredible source of support throughout a process which many joked would never end. Thanks to all of you for celebrating with me when I had papers published and commiserated lab disasters. Special thanks to Stacey, Marin and Neve Fox, for giving me the most loving home I could ask for - I couldn't list all the wonderful things they do for me everyday.

Dedication

This thesis is dedicated to my two little girls, Marin and Neve, who inspire me so much for the future, and for my grandmother Moyra Davies.

Abbreviations

AI: Atlantic Ireland

bp: Base Pair

BP: Before Present

COI: Cytochrome c Oxidase I

CNV: Copy Number Variants

DNA: Deoxyribonucleic Acid

ESU: Evolutionarily Significant Unit

ESV: Exact Sequence Variants

EU: European Union

gDNA: Genomic DNA

HWE: Hardy-Weinberg Equilibrium

ITS2: Internal Transcribed Spacer Two

ka: Kilo annum

matK: Maturase K

mRNA: Messenger RNA

MIS: Mid-Irish Sea

MPA: Marine Protected Area

MSY: Maximum Sustainable Yield

NA: Null Allele

NE: North East

N_em: Number of Effective Migrants

NGS: Next-Generation Sequencing

NMDS: Non-Metric Multidimensional Scaling

PCR: Polymerase Chain Reaction

RAD: Restriction Site Associated DNA

RAD-Seq: Restriction Site Associated DNA Sequencing

RNA: Ribonucleic Acid

SW: South West

rbcL: Ribulose Bisphosphate Carboxylase Large Chain

taq: *Thermus aquaticus*

UK: United Kingdom

USA: United States of America

UAE: United Arab Emirates

WGA: Whole Genome Amplification

Chapter 1

Thesis Introduction

1.1 Introduction

We are living through a profound period of change. Biodiversity is being directly affected by the actions of humans on a global scale, through habitat destruction and alteration, over-exploitation and global warming. Species loss is occurring at an unprecedented rate such that many consider the last two centuries to be a mass extinction event similar in magnitude to that which caused the extinction of the dinosaurs (Sodhi and Ehrlich, 2010). As the human population continues to increase, and places more pressure on natural resources, the rate of extinction of other species is expected to rise to around 1000x the normal background rate suggested by the fossil record (Frankham et al., 2004). The protection of species and ecosystems are critically important for their inherent value, their cultural significance, and for the indispensable ecosystem services they provide (Bateman et al., 2013). It is more important than ever that the myriad of natural systems governing life on Earth are understood such that detrimental effects can be minimised, and further damage prevented. Studies of ecology and conservation contribute to this understanding and provide valuable knowledge to scientists, governments, and citizens to help to protect the natural environments around us.

1.2 DNA Sequencing and Ecology

1.2.1 The Development of DNA Sequencing Technologies

The study of DNA and genetics has enabled ecologists and conservationists to develop and use powerful tools to investigate the world around them. These tools increase our understanding of the mechanics of ecosystems, and the threats they face (Frankham et al., 2004). The discipline of molecular ecology grew in tandem with the development of molecular biology techniques in the latter half of the 20th century, most notably Sanger sequencing (Sanger et al., 1977) and the polymerase chain reaction (PCR), (Mullis et al., 1986). Alongside the development of new genetic markers (Vieira et al., 2016; LaFramboise, 2009), these novel molecular methods allowed genetic variation to be directly measured, revolutionising many fields of ecology (Allan, 2010; Monsen-Collar and Dolcemascolo, 2010). The availability of

ecological genetic data meant that conservation and management decisions could increasingly be made based on empirical observations and quantifiable data (Fox et al., 2018), alongside more traditional methods of observational conservation.

At the end of the 20th century a global scientific effort was harnessed to produce the first draft human genome as part of the human genome project (Venter et al., 1998). This ground-breaking achievement involved dozens of collaborators and technicians and cost millions of dollars. By 2005, just four years after the publication of the first draft human genome, a new generation of sequencing technologies had become commercially available, capable of producing comparable amounts of sequence data to that used in the human genome project, in a matter of days (Goodwin et al., 2016). Next-generation sequencing (NGS), or high-throughput parallel sequencing, are terms used to represent this new paradigm in DNA sequencing which occurred almost 30 years after Sanger's seminal sequencing method based on chain termination was published (Sanger et al., 1977). After three decades of the dominance of Sanger's technique, new methods of determining nucleotide sequences began to emerge, which rather than rely upon molecular weight as inferred by electrophoresis, allowed the nucleotide sequence to be determined directly from the *taq*-based synthesis of double stranded DNA molecules themselves (Ronaghi et al., 1998).

The development of high-throughput, next-generation sequencing (NGS) centred around completely novel methods of library preparation and sequencing. Whilst Sanger's technique required a single highly purified DNA template, massively parallel sequencing enables the concurrent sequencing of millions of individual strands of DNA, allowing data generation on a scale many orders of magnitude greater than had previously been possible (Shendure and Ji, 2008; Williams et al., 2006).

The development and optimisation of low cost, short-read sequencing approaches have lead to dramatically decreasing sequencing costs per megabase, and therefore wider availability of the technologies and of sequence data (Goodwin et al., 2016). Whole genome sequencing is now a routine process, due to the rapid evolution of sequencing technologies and associated bioinformatics methods, with entirely new fields of molecular analysis being developed as a result of technological innovation

including RNA-seq, ChIP-seq, metabarcoding and metagenomics (Levy and Myers, 2016; Metzker, 2010; Morey et al., 2013; Reuter et al., 2015).

1.3 Applications of Next-Generation Sequencing in Ecology

Ecology is a broad-ranging subject, covering all interactions at the organismal level, and with their surrounding ecosystems. Molecular ecology is no less complex, and molecular methods have been employed right across the field of ecological study, as they have in almost every aspect of the life sciences. This thesis investigates the use of next-generation sequencing methods, and nucleotide data, to answer a range of ecological questions. Some examples of the uses of next-generation sequencing data, and how it can be used to answer ecological questions follow.

1.3.1 Genetic Management and *ex situ* Conservation.

For the management of small *ex situ* populations, such as those in zoos or aquaria, managed breeding to maintain genetic diversity and to identify evolutionarily significant units (ESU) is critical for their effective management (Frankham et al., 2004; Fox et al., 2018). Previously, population management was reliant upon observational data and the maintenance of studbooks. The inclusion of genetic data is a powerful, additional tool available to conservationists to further complement their work. Molecular methods are invaluable to the detection and limitation of inbreeding depression (Ralls et al., 1979; Lukas et al., 1994), identification of appropriate management units (Palsbøll et al., 2007) and the construction of pedigrees to aid in the recovery and management of endangered species (Haig, 1998). As ever more species become threatened, or endangered, the establishment of *ex situ* populations to protect against extinction is likely to become even more routine (Dawson et al., 2011). Development of genetic markers, such as microsatellites, will be required for the management of these species unless previously characterised markers are readily available in public databases. Shotgun sequencing, or whole-genome sequencing, allows the collection of data from across the genome of a subject (Venter et al., 1998). This non-targeted (Davey et al.,

2011) approach to DNA sequencing, allows for computational methods to be used to mine for genetic markers which can be used downstream for further analysis, including for the management of *ex-situ* populations (Castoe et al., 2015; Fox et al., 2019). The availability of genetic markers for many non-model species, and the relative ease by which markers can be developed for any species of interest, makes genetic studies much more accessible to the ecology community globally.

Genetic markers such as microsatellites or single nucleotide polymorphisms are important tools which can be used to detect gene flow, or genetic isolation between groups of individuals, of the same species (Zhang et al., 2018; Lemopoulos et al., 2019). These (potentially) genetically distinct populations, which can be identified as ESUs, may be separated due to historical isolation, reproductive isolation or may have evolved forms of adaptive isolation due to their ecosystem. Each may require separate management strategies based on both their genetic and ecological differences (Crandall et al., 2000). Excessive admixture, from the mixing of genetically or ecologically distinct populations can result in the loss of adaptation to local conditions, or reduced fitness of progeny, potentially further threatening an already at-risk species (Frankham et al., 2004).

1.3.2 Population Genetics

Population genetics, the study and quantification of genetic variation within and among populations, is concerned with the frequencies and relative fitness of genotypes within populations (Gillespie, 1998), and provides valuable information upon dispersal, the limits of population ranges and reproductive behaviour. The range of a species is influenced by many ecological factors, and the spatial genetic structure is clearly inherently linked (Janes and Batista, 2016; Jarvis et al., 2005; Tinnert et al., 2016).

Accurate identification of the genetic structure of a species enables the definition of evolutionarily significant units (Mockford et al., 2007) and management units (Palsbøll et al., 2007). Information relating to the degree of gene flow between populations is critical for effective management in order to combat damaging over-exploitation, and for the calculation of metrics such as the maximum sustainable yield (Hastings and Botsford, 2006).

There are several important factors known to influence the spatial genetic structure of natural populations. Amongst the most powerful of these are neutral processes such as genetic drift which gives rise to small random changes in genoptye frequencies caused by chance fluctuations and the loss of less frequent alleles. Non-neutral processes also act to remove deleterious or less-fit alleles and selectively promote the most beneficial allele (Swaegers et al., 2015). Migration and gene flow allow the exchange of alleles from distinct populations, reducing their genetic differentiation and homogenizing otherwise distinct populations. Mutations provide new alleles at very low frequency to the species or population, providing the raw materials for evolutionary processes and have a powerful effects upon the relative fitness of alleles and individuals (Gillespie, 1998; Baines et al., 2004; Tinnert et al., 2016; Loewe and Hill, 2010). The dynamics of a species, and sub-populations which may be linked by high-rates of gene flow, or completely genetically isolated from one another, is governed by the dynamics and interplay of these main factors. Microsatellites are sometimes assumed to be selectively neutral, in fact this is often stated as a benefit of their use as a marker for population genetics, however this is not always the case and both microsatellites and copy number variants have been discovered to be the targets of selection (Haasl and Payseur, 2013; Vychodilova et al., 2018). Population structure is influenced by selective pressures as well as non-selective. Genomic approaches which allow the genotyping of many thousands of loci from right across the target genome, allow for the incorporation of data from both selectively neutral and non-neutral markers, for example from loci associated with important traits and diseases (Sutter et al., 2007). The incorporation and comparison of data from both neutral and non-neutral markers, allows for study of the effects of selection upon the population, and its isolation from neutral processes driving population structure (Hendricks et al., 2018), which has lead in part to the development of new genetic marker types for population genetics, most notably the single nucleotide polymorphism.

Whole genome sequencing can be used for the development of novel microsatellite markers, but more recent methods have lead to the development of single-nucleotide polymorphism markers (SNPs) becoming the preferable genetic

marker for population genetics in many cases. Restriction-site associated DNA sequencing (RAD-Seq) and subsequent SNP genotyping, requires no *a priori* knowledge of the genome of the study species and can quickly generate tens of thousands of potential genetic markers. As marker number is correlated with statistical power to detect genetic structure (Coates et al., 2009), a large panel of SNPs is a powerful tool for the analysis of population structure, and to inform the management of a population.

1.3.3 Metabarcoding

Genetic, or DNA barcoding is a technique used for the identification of a species or organism, based upon the sequence of a particular genetic marker. An ideal candidate for a marker gene is one which is very highly conserved within a species, but contains nucleotide substitutions between even closely related species (Wooley et al., 2010). DNA barcoding is the colloquial name for the process where a fragment of a marker gene from an unknown organism is sequenced using one of the widely available DNA sequencing methods, and the resulting nucleotide sequence compared to database where it can be uniquely identified to a taxa (Kress et al., 2015). Several DNA barcodes exist in the literature, often specifically used for different taxa: cytochrome c oxidase I (*COI*) is commonly used for identification of animals (Hebert et al., 2003), *16S* ribosomal RNA commonly used for bacterial identification (Janda and Abbott, 2007) and the internal transcribed space (*ITS*) for fungi (Schoch et al., 2012). Several barcodes exist for plant identification including ribulose bisphosphate carboxylase large chain (*rbcL*), *ITS*, and maturase K (*matK*), (Hollingsworth et al., 2009).

Estimates of species proportions in a mixed sample are derived from the relative abundance of metabarcoding sequencing reads, however this is known to not be a reliable method of estimating relative proportions of taxa (Deagle et al., 2018; Edgar, 2017). One reason for this lack of confidence in abundance data is that different DNA barcodes are amplified from different genomes within the organism, for example, *COI* is found in the mitochondrial genome, *ITS* in the nuclear genome and *rbcL* is a chloroplastic gene. Multiple barcoding genes are present in cells, and copy number varies between species, tissue types, and physiological states (Ma and

Li, 2015; Veltri et al., 1990), giving rise to different, equally valid, descriptions of the same community, based upon variation in the frequency of each marker.

Molecular barcoding enables the characterisation of cryptic communities and has enabled ecologists to describe a wide range of microscopic bacterial and fungal communities, perform diet analysis and analyse environmental DNA (Ramirez et al., 2017; Galan et al., 2017; Hawkins et al., 2015). Metabarcoding replaced culturing based methods for bacterial analysis, which were known to miss the vast majority of diversity due to difficulties with culturing certain taxa (Riesenfeld et al., 2004). Information regarding the type and abundance of different nutrient sources is clearly of high importance for the conservation of any species and diet analysis has been a particular benefactor of the improvements in metabarcoding analysis. Understanding the dynamics and ecology of food and nutrient webs and resource foraging allows us to see the underlying processes in the functioning of ecosystems and an element of the community interactions in an ecosystem (Pompanon et al., 2012). Accurately describing forage, or diet of an organism is often reliant upon analysis of stomach products themselves, or the remains of the digestive process, usually sampled from faeces, and was previously dependant upon accurate morphological identification of material (Forin-Wiart et al., 2018). Molecular methods of taxonomic identification using high-throughput sequencing have been successfully employed to inform the conservation of several important taxonomic groups, including carnivores (Kumari et al., 2019), birds (Nota et al., 2019) and herbivores (Kartzinel et al., 2015) and pollinators (Carvell et al., 2006; Taberlet et al., 2018).

Pollen identification is a valuable resource in pollinator ecology as it allows the determination of the plant taxa providing nectar to a pollinator species (Carvell et al., 2006). Previous methods of pollen identification have been extremely labour intensive, with manual fixing, and identification of individual pollen grains to references required (Von Der Ohe et al., 2004). Comparative studies have determined that metabarcoding approaches give similar descriptions of a community, and are more sensitive to less abundant taxa, which are potentially missed using non-genetic methods (Lejzerowicz et al., 2015).

1.4 Study Species

1.4.1 *Raja undulata*

Raja undulata, the undulate skate, is an endangered species commonly found in the bycatch of commercial fishing operations around its range in the English channel and Atlantic coasts of Ireland and Portugal (Coelho et al., 2009). The species has seen accelerating decreases in numbers thought to be driven mainly by fishing pressure (Ellis et al., 2012), with the species being classified as endangered by the IUCN in 2009 (Gibson et al., 2006). With fisheries under-threat globally due to over-exploitation, and elasmobranchs known to decline with increasing fishing effort (Botsford et al., 1997), the long-term effects on this important family are unknown. In this thesis, I use the shotgun, whole genome sequencing approach discussed above to develop a novel panel of microsatellite markers for the species. These markers are then used to genotype a captive population of *R. undulata* held in multiple UK based aquaria and to assess the overall genetic health of the *ex situ* population.

Figure 1.1

1.4.2 *Apis mellifera*

Apis mellifera, the European honey bee, is one of the most commonly kept honey bees globally. It is widely recognised that many pollinators are suffering dramatic population declines globally (Department for Environment Food and Rural Affairs, 2014), and their importance to global ecosystems, and particularly their supporting role in plant and food crop pollination (Hung et al., 2018; Gallai et al., 2009), maintains their position as a priority of ecology and conservation. The study of the dynamics of the honey bee are highly applicable to inform further conservation of wild pollinator species. In this thesis, I perform metabarcoding analysis of next-generation sequence data to assess the plant community visited by bees from 15 hives, through the detection of plant DNA extracted from pollen in honey. I use three plant metabarcoding markers in parallel to investigate some of the biases inherent in the library preparation processes associated with metabarcoding by next-generation sequencing.

Figure 1.2

1.4.3 *Homarus gammarus*

The European lobster (*Homarus gammarus*) is a marine decapod found on rocky coastline throughout Northern Europe and the Mediterranean. It is valued as a prized seafood and as a result has been the focus of intense fishing pressure for generations (Elliott and Holden, 2017; Ingebrigt et al., 2005), with precipitous declines recorded in many regions (Kleiven et al., 2018). Protective measures and legislation have been established in many countries with *H. gammarus* fisheries to protect this precious resource. Information regarding the population dynamics of this species, with its long larval mode of dispersal, is vital to better inform its conservation. In this thesis, I perform an assessment of the population structure of the species in samples from several fisheries around the coasts of the UK and Ireland. As well as providing important information regarding genetic structure of *H. gammarus*, I also perform parallel investigations using microsatellite markers and high-throughput SNP markers to assess the suitability and practicality of each for population genetics in this species.

Figure 1.3

1.5 Thesis Aims and Chapters

The aim of this thesis is to use molecular and bioinformatics techniques, based around next-generation sequence data, to answer a range of ecological questions. I show the benefit of the inclusion of genetic data to ecology and conservation, and the wide-ranging influence which the NGS revolution has had on our field.

An outline of the specific aims of the thesis follows:

(A) To optimise novel microsatellite markers in *Raja undulata*, and use them to assess the genetic health of a small, *ex situ* population.

(B) To develop, implement and demonstrate a novel methodology to develop new microsatellite markers using multiple genomic datasets.

(C) To use genetic markers to investigate the population structure of the commercially important *Homarus gammarus* fisheries of the UK and Ireland.

(D) To perform a comparative study into effective genetic markers for the analysis of the population genetics of *Homarus gammarus*.

(E) To use metabarcoding techniques to analyse the plant forage of *Apis mellifera* hives, and to investigate the bias in several plant metabarcoding markers.

This thesis is composed of four data chapters:

In Chapter two, I build the case for inclusion of genetic data in the management of *ex situ* populations and develop and implement a novel panel of microsatellite markers derived from NGS data.

In Chapter three, I develop a novel approach to microsatellite panel design from NGS data which incorporates multiple genomic data sets to improve the rate of successful microsatellite marker development.

In Chapter four, I perform parallel investigations into the population genetics of the commercially important European lobster, highlighting the relative power and application of traditional microsatellites and high-throughput SNP genotyping methods for population genetics.

In Chapter five, I analyse parallel data sets of plant metabarcoding data to investigate biases inherent in marker choice. I assess whether multiple barcoding markers offer any additional power to assign taxonomies and evaluate confidence in community descriptions.

References

Allan, G. (2010). Molecular genetic techniques and markers for ecological research. *Nature Education Knowledge*, 3(10):2.

Baines, F., Das, A., Mousset, S., and Stephan, W. (2004). The role of natural selection in genetic differentiation of worldwide populations of *Drosophila ananassae*. *Genetics*, 168(4):1987–98.

Bateman, I., Harwood, A., Mace, G., Watson, R., Abson, D., Andrews, B., Binner, A., Crowe, A., Day, B., Dugdale, S., Fezzi, C., Foden, J., Hadley, D., Haines-Young, R., Hulme, M., Kontoleon, A., Lovett, A., Munday, P., Pascual, U., Paterson, J., Perino, G., Sen, A., Siriwardena, G., van Soest, D., and Termansen, M. (2013). Bring ecosystem services into economic decision-making: Land use in the United Kingdom. *Science*, 341(6141):45–50.

Botsford, L., Castilla, J., and Peterson, C. (1997). The management of fisheries and marine ecosystems. *Science*, 277(5325):509–515.

Carvell, C., Roy, D., Smart, S., Pywell, R., Preston, C., and Goulson, D. (2006). Declines in forage availability for bumblebees at a national scale. *Biological Conservation*, 132(4):481 – 489.

Castoe, T., Poole, A., de Koning, A., Jones, K., Tomback, D., Oyler-McCance, S., Fike, J., Lance, S., Streicher, J., Smith, E., and Pollock, D. (2015). Correction: Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PLOS ONE*, 7(2):e30953.

Coates, B., Sumerford, D., Miller, N., and Kim, K. (2009). Comparative performance of single nucleotide polymorphism and microsatellite markers for population genetic analysis. *Journal of Heredity*, 100(5):556–564.

Coelho, R., Bertozzi, M., Ungaro, N., and Ellis, J. (2009). *Raja undulata.* the IUCN red list of threatened species 2009: e.t161425a5420694.

Crandall, K., Bininda-Edmonds, O., Mace, G., and Wayne, R. (2000). Considering evolutionary processes in conservation biology: an alternative to 'evolutionarily significant units'. *Trends in Ecology and Evolution*, 15(7):290–295.

Davey, J., Hohenlohe, P., Etter, P., Boone, J., Catchen, J., and Blaxter, M. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12(7):499–510.

Dawson, T., Jackson, S., House, J., Prentice, I., and Mace, G. (2011). Beyond predictions: Biodiversity conservation in a changing climate. *Science*, 332(6025):52–58.

Deagle, B., Thomas, A., McInees, J., Clarke, L., Vesterinen, E., Clare, E., Kartzinel, T., and Eveson, J. (2018). Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? *Molecular Ecology*, 28(2):391–406.

Department for Environment Food and Rural Affairs (2014). The national pollinator strategy: for bees and other pollinators in England. United Kingdom.

Edgar, R. (2017). Unbias: An attempt to correct abundance in 16S sequencing, with limited success. *bioRxiv*, 10.1101/124149.

Elliott, M. and Holden, J. (2017). UK Sea Fisheries Statistics 2017. (Office of National Statistics: Marine Management Organisation).

Ellis, J., McCully, S., and Brown, M. (2012). An overview of the biology and status of undulate ray *Raja undulata* in the North-east Atlantic ocean. *Journal of Fish Biology*, 80(5):1057–74.

Forin-Wiart, M., Poulle, M., Piry, S., Cosson, J., Larose, C., and Galan, M. (2018). Evaluating metabarcoding to analyse diet composition of species foraging in anthropogenic landscapes using Ion Torrent and Illumina sequencing. *Scientific Reports*, 8(3):474–489.

Fox, G., Darolti, I., Hibbitt, J., Preziosi, R., Fitzpatrick, J., and Rowntree, J. (2018). Genetic assessment of *ex situ* populations to aid species conservation and maintain heterozygosity in non-model species. *Journal of Zoo and Aquarium Research*, 6(2):50–56.

Fox, G., Preziosi, R., Antwis, R., Benavides-Serrato, M., Combe, F., Harris, W., Hartley, I., de Kort, S., Nekaris, A., and Rowntree, J. (2019). Multi-individual Microsatellite identification: a multiple genome approach to microsatellite design (MiMi). *Molecular Ecology Resources*, 19(6):1672–1680.

Frankham, R., Ballou, J., and Briscoe, D. (2004). *A Primer of Conservation Genetics*. Cambridge University Press, Cambridge, USA.

Galan, M., Pons, J., Tournayre, O., Pierre, E., Leuchtmann, M., Pontier, D., and Charbonnel, N. (2017). Metabarcoding for the parallel identification of several hundred predators and their prey: Application to bat species diet analysis. *Molecular Ecology Resources*, 18(3):474–489.

Gallai, N., Salles, J., Settele, J., and Vassiere, B. (2009). Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. *Ecological Economics*, 68(3):810– 821.

Gibson, C., Valenti, S., Fowler, S., and Fordham, S. (2006). The conservation status of northeast Atlantic chondrichthyans; report of the IUCN shark specialist group northeast Atlantic regional red list workshop. *VIII + 76pp. IUCN SSC Shark Specialist Group*.

Gillespie, J. (1998). *Population Genetics. A Concise Guide*. The Johns Hopkins University Press, The John Hopkins Press Ltd., London.

Goodwin, S., McPherson, J., and McCombie, W. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351.

Haasl, R. and Payseur, B. (2013). Microsatellites as targets of natural selection. *Molecular Biology and Evolution*, 30(2):285–298.

Haig, S. M. (1998). Molecular contributions to conservation. *Ecology*, 79(2):413–425.

Hastings, A. and Botsford, L. (2006). Persistence of spatial populations depends on returning home. *PNAS*, 103(15):6067–6072.

Hawkins, J., de Vere, N., Griffith, A., Ford, C., Allainguillaume, J., Hegarty, M., and Baillie L, Adams-Groom, B. (2015). Using DNA metabarcoding to identify the floral composition of honey: A new tool for investigating honey bee foraging preferences. *PLOS ONE*, 10(8):e0134735.

Hebert, P., Cywinska, A., Ball, S., and deWaard, J. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B*, 270(1512):313–21.

Hendricks, S., Anderson, E., Antao, T., Bernatchez, L., Forester, B., Garner, B., Hand, B., Hohenlohe, P., Kardos, M., Koop, B., Sethuraman, A., Waples, R., and Luikart, G. (2018). Recent advances in conservation and population genomics data analysis. *Evolutionary Applications*, 11(8):1197–1211.

Hollingsworth, P., Forrest, L., Spouge, J., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M., Cowan, R., Erickson, D., Fazekas, A., Graham, S., James, K., Kim, K., Kress, W. J., Schneider, H., van AlphenStahl, J., Barrett, S., van den Berg, C., Bogarin, D., Burgess, K., Cameron, K., Carine, M., Chacón, J., Clark, A., Clarkson, J., Conrad, F., Devey, D. S., Ford, C., Hedderson, T., Hollingsworth, M., Husband, B., Kelly, L., Kesanakurti, P., Kim, J., Kim, Y., Lahaye, R., Lee, H., Long, D., Madriñán, S., Maurin, O., Meusnier, I., Newmaster, S., Park, C., Percy, D., Petersen, G., Richardson, J., Salazar, G., Savolainen, V., Seberg, O., Wilkinson, M., Yi, D., and Little, D. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, 106(31):12794–12797.

Hung, K., Kingston, J., Albrecht, M., D.A., H., and Kohn, J. (2018). The worldwide importance of honey bees as pollinators in natural habitats. *Proceedings of the Royal Society B*, 285(20172140).

Ingebrigt, U., Benavente, G., and Browne, R. (2005). A regional development strategy for stock enhancement of clawed lobsters (*Homarus gammarus*): Development of juvenile lobster production methodologies. Norwegian Institute for Nature Research(NINA Report).

Janda, J. and Abbott, S. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *Journal of Clinical Microbiology*, 45(9):2761–2764.

Janes, J. and Batista, P. (2016). Chapter two - the role of population genetic structure in understanding and managing pine beetles. *Advances in Insect Physiology*, 50:75–100.

Jarvis, A., Yeaman, S., Guarino, L., and Tohme, J. (2005). The role of geographic analysis in locating, understanding, and using plant genetic diversity. *Methods in Enzymology*, 395:279–298.

Kartzinel, T., Chen, P., Coverdale, T., Erickson, D., Kress, W., Kuzmina, M., Rubenstein, D., Wang, W., and Pringle, R. (2015). DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proceedings of the National Academy of Sciences of the United States of America*, 112(26):8019–8024.

Kleiven, A., Moland, E., Olsen, E., and Knutsen, J. (2018). *Integrated Coastal Zone Management*, chapter Lobster Reserves in Coastal Skagerrak – An Integrated Analysis of the Implementation Process. Springer Vieweg.

Kress, J., García-Robledo, C., Uriarte, M., and Erickson, D. (2015). DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology and Evolution*, 30(1):25–35.

Kumari, P., Dong, K., Eo, K., Lee, W., Kimura, J., and Yamamoto, N. (2019). DNA metabarcoding-based diet survey for the Eurasian otter (*Lutra lutra*): Development of a Eurasian otter-specific blocking oligonucleotide for 12S *rRNA* gene sequencing for vertebrates. *PLOS ONE*, 14(12):e0226253.

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res*, 37(13):4181–4193.

Lejzerowicz, F., Esling, P., Pillet, L., Wilding, T., Black, K., and Pawlowski, J.

(2015). High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. *Scientific Reports*, 5(Article 13932).

Lemopoulos, A., Prokkola, J., Uusi-Heikkilä, S., Vasemägi, A., Huusko, A., Hyvärinen, P., Koljonen, M., Koskiniemi, J., and Vainikka, A. (2019). Comparing RADseq and microsatellites for estimating genetic diversity and relatedness — Implications for brown trout conservation. *Ecology and Evolution*, 9(4):2106–2120.

Levy, S. and Myers, R. (2016). Advancements in Next-Generation Sequencing. *The Annual Review of Genomics and Human Genetics*, 17:95–115.

Loewe, L. and Hill, W. (2010). The population genetics of mutations: good, bad and indifferent. *Philos Trans R Soc Lond B Biol Sci*, 27(365):1153–1167.

Lukas, F., Keller, P., Smith, J., Hochachka, W. M., and Stearns, S. (1994). Selection against inbred song sparrows during a natural population bottleneck. *Nature*, 372:356–357.

Ma, J. and Li, X. (2015). Organellar genome copy number variation and integrity during moderate maturation of roots and leaves of maize seedlings. *Current Genetics*, 61(4):591–600.

Metzker, M. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11:31–46.

Mockford, S., Herman, T., Snyder, M., and Wringht, J. (2007). Conservation genetics of Blanding's turtle and its application in the identification of evolutionarily significant units. *Conservation Genetics*, 8(1):209–219.

Monsen-Collar, K. J. and Dolcemascolo, P. (2010). Using molecular techniques to answer ecological questions. *Nature Education Knowledge*, 3(10):1.

Morey, M., Fernández-Marmiesse, A., Castiñeiras, D., Fraga, J., Couce, M., and Cocho, J. (2013). A glimpse into past, present, and future DNA sequencing. *Molecular Genetics and Metabolism*, 110(1-2):3–24.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. (1986). Specific enzymatic amplification of DNA *In Vitro*: The polymerase chain reaction. *Cold Spring Harbour Symposium on Quantitative Biology*, 51(1):263–273.

Nota, K., Downing, S., and Iyengar, A. (2019). Metabarcoding-based dietary analysis of hen harrier (*Circus cyaneus* in Great Britain using buccal swabs from chicks). *Conservation Genetics*, 20(6):1389–1404.

Palsbøll, P., Bérubé, M., and Allendorf, F. (2007). Identification of management units using population genetic data. *Trends in Ecology & Evolution*, 22(1):11–16.

Pompanon, F., Deagle, B., Symondson, W., Brown, D., Jarman, S., and Tablerlet, P. (2012). Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology*, 21(8):1931–1950.

Ralls, K., Brugger, K., and Ballou, J. (1979). Inbreeding and juvenile mortality in small populations of ungulates. *Science*, 206(4422):1101–1103.

Ramirez, K., Knight, C., de Hollander, M., Brearley, F., Constantinides, B., Cotton, A., Creer, S., Crowther, T., Davison, J., Delgado-Baquerizo, M., Dorrepaal, E., Elliott, D., Fox, G., Griffiths, R., Hale, C., Hartman, K., Houlden, A., Jones, D., Krab, E., Maestre, F., McGuire, K., Monteux, S., Orr, C., van der Putten, W., Roberts, I., Robinson, D., Rocca, J., Rowntree, J., Schlaeppi, K., Shepherd, M., Singh, B., Straathof, A., Bhatnagar, J., Thion, C., van der Heijden M.G.A., and de Vries, F. (2017). Detecting macroecological patterns in bacterial communities across independent studies of global soils. *Nature Microbiology*, 3:189–196.

Reuter, J., Spacek, D., and Snyder, M. (2015). High-throughput sequencing technologies. *Molecular Cell*, 58(4):586–597.

Riesenfeld, C., Schloss, P., and Handelsman, J. (2004). Metagenomics: Genomic analysis of microbial communities. *Annual Review of Genetics*, 38:525–552.

Ronaghi, M., Uhlen, M., and Nyrén, P. (1998). A sequencing method based on real-time pyrophosphate. *Science*, 281(5375):363–365.

Sanger, F., Nicklen, S., and Coulson, A. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*, 74(12):5463–5467.

Schoch, C., Seifert, K., Huhndorf, S., Robert, V., Spouge, J., Levesque, A., and Chen, W. (2012). Nuclear ribosomal internal transcribed spacer ITS region as a universal DNA barcode marker for fungi. *PNAS*, 109(16):6241–6246.

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145.

Sodhi, N. and Ehrlich, P. (2010). *Conservation Biology for All*. Oxford University Press, Great Clarendon Street, Oxford, UK.

Sutter, N., Bustamante, C., Chase, K., Gray, M., Zhao, K., and Zhu, L. (2007). A single IGF1 allele is a major determinant of small size in dogs. *Science*, 316(5821):112–115.

Swaegers, J., Mergeay, J., Van Geystelen, A., Therry, L., Larmuseau, M., and Stoks, R. (2015). Neutral and adaptive genomic signatures of rapid poleward range expansion. *Molecular Ecology*, 24:6163–6176.

Taberlet, P., Bonin, A., Zinger, L., and Coissac, E. (2018). *Environmental DNA*. Oxford University Press, Great Clarendon Street, Oxford, OX2 6DP.

Tinnert, J., Hellgren, O., Lindberg, J., Koch-Schmidt, P., and Forsman, A. (2016). Population genetic structure, differentiation, and diversity in *Tetrix subulata* pygmy grasshoppers: roles of population size and immigration. *Ecology and Evolution*, 6(21):7831–7846.

Veltri, K., Espiritu, M., and Singh, G. (1990). Distinct genomic copy number in mitochondria of different mammalian organs. *Journal of Cellular Physiology*, 143(1):160–164.

Venter, J., Adams, M., Sutton, G., Kerlavage, A., Smith, H., and Hunkapiller, M. (1998). Shotgun sequencing of the human genome. *Science*, 280(5369):1540–1542.

Vieira, M., Santini, L., Diniz, A., and Munhoz, C. (2016). Microsatellite markers: what they mean and why they are so useful. *Genetics and Molecular Biology*, 39(3):312–328.

Von Der Ohe, W., Persano Oddo, L., Piana, M., Morlot, M., and Martin, P. (2004). Harmonized methods of melissopalynology. *Apidologie*, 35:S18–S25.

Vychodilova, L., Necesankova, M., Albrechtova, K., Hlavac, J., Modry, D., Janova, E., Vyskocil, M., Mihalca, A., Kennedy, L., and Horin, P. (2018). Genetic diversity and population structure of African village dogs based on microsatellite and immunity-related molecular markers. *PLOS ONE*, 13(6):e0199506.

Williams, R., Peisajovich, S., Miller, O., Magdassi, S., Tawfik, D., and Griffiths, A. (2006). Amplification of complex gene libraries by emulsion PCR. *Nature Methods*, 3(7):545–550.

Wooley, J., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLOS ONE*, 6(2):e1000667.

Zhang, H., Li, H., and Li, Y. (2018). Identifying evolutionarily significant units for conservation of the endangered *Malus sieversii* using genome-wide RADseq data. *Nordic Journal of Botany*, 36(7).

Chapter 2

Application of Genetic Data to *ex situ* Conservation.

2.1 Genetic assessment of *ex situ* populations to aid species conservation and maintain heterozygosity in non-model species.

Graeme Fox,[1] Iulia Darolti,[2] Jean-Denis Hibbitt,[3] Richard F. Preziosi,[1] John F. Fitzpatrick,[4] Jennifer K. Rowntree,[1]

[1]Ecology and Environment Research Centre, Department of Natural Sciences, Manchester Metropolitan University, Chester Street, Manchester, United Kingdom, M1 5GD

[2]Department of Genetics, Evolution and Environment, University College London, London, United Kingdom, WC1E 6BT

[3]SEA LIFE Global, Merlin Animal Welfare and Development, Lodmoor County Park, Weymouth SEA LIFE Adventure Park, Preston Road, Weymouth, Dorset, United Kingdom, DT4 7SX

[4]Department of Zoology/Ethology, Stockholm University, Stockholm, Sweden, SE-106 91

Key words: Raja undulata · Microsatellite markers · Population genetic structure · Elasmobranchii · Next-generation sequencing

Author contributions: GF, JDH, RFP, JFF and JKR conceived and designed the project; ID and JDH collected the samples; GF performed the lab work; GF performed the data analysis; GF wrote the chapter.

2.2 Publication Reference

This chapter is published at:

Fox, G., Darolti, I., Hibbitt, J.D., Preziosi, R.F., Fitzpatrick, J.L. and Rowntree, J.K. (2018) Genetic assessment of *ex situ* populations to aid species conservation and maintain heterozygosity in non-model species. *Journal of Zoo & Aquarium Research* 6(2). pp. 50-56.

JZAR JOURNAL OF ZOO AND AQUARIUM RESEARCH
www.jzar.org

EAZA

Research article

Bespoke markers for ex-situ conservation: application, analysis and challenges in the assessment of a population of endangered undulate rays

Graeme Fox[1,2], Iulia Darolti[3,2], Jean-Denis Hibbitt[4], Richard F. Preziosi[1,2], John L. Fitzpatrick[5,2], Jennifer K. Rowntree[1,2]

[1]School of Science and the Environment, Manchester Metropolitan University, John Dalton East, Manchester, M1 5GD, UK
[2]Faculty of Life Sciences, The University of Manchester, Manchester, M13 9PT, UK
[3]Department of Genetics, Evolution and Environment, University College London, London, WC1E 6BT, UK
[4]SEA LIFE Global, Merlin Animal Welfare and Development, Lodmoor Country Park, Weymouth SEA LIFE Adventure Park, Preston Road, Weymouth, Dorset, DT4 7SX, UK.
[5]Department of Zoology/Ethology, Stockholm University, Stockholm, SE-106 91, Sweden

Correspondence: Dr Jennifer K. Rowntree; j.rowntree@mmu.ac.uk

Figure 2.1

2.3 Abstract

Genetic data are important and informative in the management of *ex situ* populations. Where the risk of inbreeding is particularly great, it is critical that tools are employed that allow for the quantification of genetic variation and to identify potential breeding pairs. This study demonstrates the rapid application of laboratory and bioinformatics techniques to develop a novel microsatellite marker panel for use with a population of the endangered undulate ray (*Raja undulata*) and shows how a minimally invasive sampling method can be used with aquarium-dwelling individuals. The study assesses the population and investigates how informative a small microsatellite marker panel is to the conservation of a restricted *ex situ* group. It was found that after a single captive generation of *R. undulata* there is no detectable evidence of reduced heterozygosity and no observable aquaria effects or differences between the generations. In conclusion, the study demonstrates that it is practical, quick and informative to develop a bespoke panel of markers to aid *ex situ* conservation efforts of non-model species and make recommendations that these processes should constitute the minimum effort required in managing such a population.

2.4 Introduction

The elasmobranchii are a subclass of carnivorous, cartilaginous fish, including the sharks, rays, skates and sawfish. These species are found extensively in coastal, demersal and pelagic marine habitats and an additional minority inhabit freshwater systems (Compagno, 1990). Common traits include slow growth and low productivity (Frisk et al., 2001; Walker and Hislop, 1998), resulting in high vulnerability and slow response to over-exploitation from fishing activities (Ferretti et al., 2010; Smith et al., 1998). Recorded declines in elasmobranch populations over recent decades are typically associated with increasing fishing effort; an effect which can be seen in oceans the world over, for example in the Gulf of Mexico (Shepherd and Myers, 2005); the northwest Atlantic (Baum et al., 2003); the Mediterranean Sea (Ferretti et al., 2008); the Sea of Japan (Nakano, 1999) and the Indian Ocean (Appukuttan and Nair, 1988). Whether fishing effort targets elasmobranchs specifically (Rose, 1998; Stevens et al., 2000) or they are a common feature of bycatch (Oliver et al., 2015), with the majority of global fisheries at risk of over-exploitation (Botsford et al., 1997) the long-term effect on

elasmobranch populations is largely unknown (Baum et al., 2003). The undulate ray (*Raja undulata*) is an endangered skate often present in bycatch of commercial trawl fishing operations off the south coast of England, France, western Ireland and southern Portugal (Coelho et al., 2009). Existing in small isolated populations, the species has recorded declines of up to 80% in some areas since the early 1980s, which has been directly attributed to fishing activities (Ellis et al., 2012). In 2009, the species was classified as endangered by the IUCN (Gibson et al., 2006). A managed breeding and monitoring program (Mon-P) was established in 2010 by the European Association of Zoos and Aquaria (EAZA) in response to the new IUCN classification and a European Union ban on the landing of this skate species was put in place. Currently, 36 aquaria across nine countries hold *R. undulata*. As part of the larger European breeding program, a small captive group is maintained across several UK aquaria, comprising a mixture of wild-caught and captive-bred individuals. Very little is known about the genetic diversity or population genetic structure of this species either in captivity or in the wild. The elasmobranchii are a charismatic focal point of interest for the general public in aquaria and are the subject of intense conservation effort to manage their *ex situ* conservation. With >100 chondrichthyan species present in European zoos and aquaria (8.6% of all known elasmobranch species), there is great interest in the community for methods and techniques for sustainable conservation of these animals (Janse et al., 2017). Non-random mating and genetic drift are major concerns for small populations and can have devastating implications for the evolutionary potential of the group. The small size of the population limits potential reproductive pairings, and inbreeding becomes a risk with the increased probability of a pair of individuals being related to one another (Witzenberger and Hochkirch, 2011). Prolonged inbreeding in a closed population increases the probability of progeny being homozygous at a given locus, resulting in the overall reduction of heterozygosity of the group after successive generations. Genetic drift and adaptation to captivity can also contribute to the loss of rare alleles and overall reduction in heterozygosity (Price and Hadfield, 2013; Willoughby et al., 2014). It is widely recognised that the fitness of a population is inversely related to allelic homozygosity, and severe effects, such as loss of viability or infertility, can present after just a few generations of close inbreeding (Frankham et al., 2004). These detrimental effects are cumulative as they are amplified by successive generations in captivity (Christie et al., 2012). As a result, the longer it has been in isolation, the less-well suited a captive population becomes to providing individuals for release

(Earnhardt, 2010; Lacy, 2012). It is imperative, therefore, that the genetic variation present at the founding of the *ex situ* population be carefully retained and inbreeding avoided through strategic genetic management of the population (Fernández et al., 2004; Pelletier et al., 2009) Under ideal conditions, during the establishment of a new *ex situ* population, the entire group should be assessed using genetic markers to estimate the diversity of the cohort and help establish a baseline of genetic diversity, to identify any genetic similarity of founding individuals and to support future management. In the case of an existing population, genetic markers should be used even in the presence of detailed keeper reports and pedigrees; whilst these resources contain valuable information, they are limited in scope to the time that the individuals (or their ancestors) have been known to the relevant managers. The most common genetic marker used in analyses of this type is the microsatellite; short, repetitive, hypervariable regions of DNA that appear to be a feature universal to all genomes. Microsatellite marker panels are available in online databases for many species and published, optimised methodologies are available for developing novel sets of markers (Castoe et al., 2015; Griffiths et al., 2016). As the rate of species extinction is elevated above the background rate (Pimm et al., 2014) and there is potential for an unprecedented increase in the number of *ex situ* populations being managed across a wide range of taxa (Dawson et al., 2011), it is imperative that general best practice guidelines in genetic management are established now. In line with published recommendations (Witzenberger and Hochkirch, 2011; Janse et al., 2017), the current best practice is argued to be the use genetic markers to characterise the diversity and relatedness of individuals in a captive breeding program and this should be the minimum standard required for the establishment, or maintenance, of any *ex situ* conservation program. When sampling for the collection of DNA, the aim should be to minimise stress or discomfort experienced by the subject whilst collecting high quality genomic template, especially in the case of an endangered or threatened species. Tissue sampling or destructive biopsy is clearly counterproductive in some cases, therefore the development and testing of non- or minimally invasive sampling methods is paramount. Here, a minimally invasive sampling method, developed for wild elasmobranchs (Lieber et al., 2013), is tested on aquarium specimens and found to be highly successful when combined with an off-the-shelf DNA extraction kit that enables isolation of high-purity DNA from the mucus layer. In this investigation, bioinformatics techniques are used to develop a novel microsatellite marker panel suitable for use in *Raja undulata*, using Illumina shotgun next-generation

sequencing data. These markers are then optimised in the laboratory and used to characterise a small *ex situ* population. The viability and confidence with which the small marker panel can be used for population management is assessed, whilst providing a snapshot of the diversity contained within this population of captive elasmobranchs.

2.5 Materials and Methods

2.5.1 Microsatellite Marker Development

High-throughput, shotgun genomic sequencing can be used in order to identify microsatellite regions in the target genome. High quality, large molecular weight, genomic DNA is essential for successful next-generation sequencing and can be collected in a variety of ways, often using a species-specific method. Samples of blood, tissue or buccal swabs (Dunn et al., 2010) are also commonly used for genetic sampling. In this instance, tissue samples were obtained from a female ray that had been euthanised due to terminal ill health resulting from a severe fungal infection of the lateral line system. A range of tissue samples were taken from the animal post euthanasia under the guidance of Mark F. Stidworthy, veterinary pathologist at International Zoo Veterinary Group (IZVG). DNA was extracted from 25 mg heart tissue using the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany), following the manufacturer's protocol and checked for quality on a NanoDrop ND-1000 spectrophotometer (260/280 >1.4) and on a 1% agarose electrophoresis gel. A sequencing library was prepared using 50ng genomic DNA and analysed on an Illumina MiSeq platform at the University of Manchester (UK) Genomics Facility using a shotgun, paired-end 2*250 sequencing methodology (Nextera DNA Library Preparation Kit, Illumina, San Diego, USA). In total, 11,019,590 raw sequencing reads were produced from the MiSeq run. Low quality regions were removed from each end of the reads, reads were trimmed using the average quality score over a sliding-window of 4nt and a quality threshold of 20, and a minimum length of 50nt was applied using Trimmomatic v0.0.4 (Bolger et al., 2014). If either of the paired-end reads failed a quality check, both reads were discarded, thus maintaining parity in the paired-end data. A majority (92%) of reads successfully passed quality filtering and were subsequently screened for potential microsatellite loci using pal_finder v0.02.04 software (Castoe et al., 2015). Non-perfect repeat loci were discarded and a minimum motif size of 3nt was implemented (Griffiths et al., 2016). Primer sequences were designed using

Primer3 v.4.0.0 (Koressaar and Remm, 2007; Untergasser et al., 2012) using conditions optimised for the Qiagen Type-it microsatellite PCR kit (Qiagen, Hilden, Germany) (optimum length: 25nt, minimum length: 18nt, maximum length: 30nt, minimum GC%: 45%, maximum GC%: 65%, minimum melting temperature: 62°C, maximum melting temperature: 75°C, optimum melting temperature: 68°C, with remaining options set to the Primer3 default values); a set of PCR reagents designed specifically for amplification of microsatellite loci. The pal_finder process produced 698 potential loci that were ranked by predicted utility as a microsatellite marker (larger motifs preferred) and the primer sequences from the first 24 results were used to purchase DNA oligos from Sigma Aldrich (Missouri, USA) (scale: 0.025 µmole, purification: DST).

2.5.2 Sampling

For characterisation of the microsatellite loci, the 35 captive *R. undulata* (17 wild-caught, 18 captive-bred) were sampled using a modified form of the minimally-invasive sampling method developed for wild elasmobranch sampling by Lieber and colleagues (Lieber et al., 2013), a method not known to have been previously demonstrated on captive animals. Small (1.5 cm x 2.5 cm), autoclaved sections of kitchen scouring pad (Vale Mill Ltd., Rochdale) were used to gently scrub the pectoral fin of the rays against the direction of the scales removing epidermal mucus secretions. Inter-species contamination was controlled, to the best of our ability, through the use of the species-specific PCR primers. As the markers were designed in a sample taken from excised heart tissue of an undulate ray (low risk of contamination), successful marker amplification implies a lack of contamination as the target DNA was of the same taxa as the heart sample. Intraspecies contamination is more difficult to control for; however, it appears not to have been an issue, as microsatellite peak traces did not show multiple banding. The pads were immediately placed into individual tubes of absolute ethanol and stored at -80°C. During DNA extraction, extraneous pad was removed and DNA was extracted using the E.Z.N.A. Mollusc DNA Kit (Omega Bio-Tek, Norcross, USA); the use of chloroform:isoamyl alcohol (24:1) successfully isolating the mucus, precipitating proteins and producing high quality DNA extract. Elution was performed in 100 µL MilliQ water and used in downstream PCR for genotyping. This sampling technique reduces stress and damage to the animal as it minimises, or eliminates in some cases, the time the specimen spends out of the water during sampling. The technique could potentially be applicable to any captive elasmobranch with a mucus layer on the skin. A total of 35 animals were sampled from 10

different aquaria (Table 2.1). Samples were also taken from several related *Raja* species (*R. microcellata, R. brachyura, R. montagui and R. clavata*) in order to test the cross-compatibility of the primers.

2.5.3 Marker Amplification

Twenty-four potential markers were tested in the laboratory, of which eight successfully amplified. PCR amplifications of 5µL total volume were performed using the Qiagen Type-it Microsatellite PCR Kit (Qiagen, Hilden, Germany). Reactions consisted of 2.5µL Type-it mastermix, 1.5µL PCR grade H2O, 0.5µL genomic DNA at 20ng/µL and 0.5µL primer pair at 2µM. This 5µL reaction was amplified under the conditions specified by the PCR kit (5 min 95°C, 28x 30 sec 95°C, 90 sec 60°C, 30 sec 72°C, 30 min 60°C) and successful amplifications were confirmed by the presence of bands on a 1% agarose electrophoresis gel. A three-primer universal-tailed approach was used to label amplicons with fluorescent moieties (Blacket et al., 2012) and fragment length reported using an Applied Biosystems 3730 DNA analyser capillary sequencer (Applied Biosystems, Foster City, California, USA) and GeneScan 500 LIZ dye size standard (Thermo Fisher Scientific, Carlsbad, USA) at the University of Manchester DNA Sequencing Facility.

2.5.4 Population Genetic Analysis

Raw data analysis was performed using GeneMapper 5.0 (Thermo Fisher Scientific, Carlsbad, USA) and confirmed that loci were scoreable and polymorphic. The novel markers were analysed for evidence of linkage disequilibrium and for deviation from Hardy–Weinberg equilibrium using GenePop v.4.2 online (Raymond and Rousset, 1995; Rousset, 2008). Estimates of pairwise relatedness were calculated for every pair of individuals using the triadic likelihood estimator of relatedness, a measure suited to a relatively small number of markers, implemented in "Coancestry" using the R (R Development Core Team, 2008) package "related" (Pew et al., 2015). The rate of heterozygosity, inbreeding coefficient and measures of genetic distance were calculated using the "adegenet" package in R (Jombart, 2008; Jombart and Ahmed, 2011; Rogers, 1972). Rates of allelic richness and private alleles were identified using the R package "PopGenReport" (Adamack and Gruber, 2014). The data were split by generation, and comparisons were drawn between each generation. In this instance, all wild-caught individuals were compared to all captive-bred offspring, as at the time of sampling there

was only a single generation captive population (F1 generation).

2.6 Results

Eight polymorphic microsatellite markers were initially characterised and every marker demonstrated to amplify consistently at an annealing temperature of 60°C, advantageous for multiplex PCR. These novel markers were used to genotype 35 captive *R. undulata* individuals at the eight loci. GENEPOP results for linkage disequilibrium (LD) showed that 48% of total marker pairs exhibited significant evidence of LD; however, when just the wild-caught individuals were tested, this percentage was reduced to zero. GENEPOP was also used to check for deviation from the expected allele frequencies of Hardy–Weinberg. Three markers showed significant deviation in the total population and a single marker (Ru13) showed deviation from expected frequencies in the wild-caught animals only. This marker (Ru13) was subsequently removed from the analysis. Summary statistics for this and the remaining seven markers are given below in Table 2.2.

A success rate of 98% was achieved in obtaining genotypic data. Average allelic richness was 7.0 in the wild-caught group, 6.4 in the captive-bred group and 1.7 per aquarium. The average observed rate of heterozygosity at each marker was 0.81. Observed heterozygosity (H_{OBS}) and the average estimated inbreeding coefficient (r) were calculated for the wild-caught animals (H_{OBS}=0.80, r=0.21±0.003) and the first generation, captive-bred individuals (H_{OBS}=0.83, r=0.18±0.005).

There was no significant difference in either heterozygosity (two sample t-test, t=0.52644, df=10.171, P=0.6099) or the average inbreeding coefficient (two sample t-test: t=-1.0356, df=14.225, p=0.3177) between wild-caught and captive-bred individuals. One to three private alleles were discovered in six of the 10 aquaria (aquarium population size ranging from 1–9 individuals). A nonmetric multidimensional scaling (NMDS) analysis of Provesti's genetic distance among individuals (Figure 2.2), calculated using the R (R Development Core Team, 2008) package "vegan" (Oksanen et al., 2017), provides a visual interpretation of the genetic similarity of individuals. The calculated stress value of the NMDS was 0.17, the lowest stress value of each of the measures of genetic distance calculated using the "adegenet" (Jombart, 2008; Jombart and Ahmed, 2011) package in R. A stress value of <0.2 indicates a fair fit of the data in the NMDS analysis (Kruskal, 1964).

The minimally invasive extraction method and the seven primer pairs were tested with

several other species of the *Raja* genus (species listed previously) and were demonstrated to successfully amplify polymorphic loci in every species tested, suggesting good cross-species compatibility of the primers and sampling technique. Allelic range in these species very closely matched those discovered in *R. undulata*, (Table 2.3). Four or fewer samples from each species were tested and, therefore, more detailed locus statistics are not provided here.

2.7 Discussion

The goal of this study was to develop and optimise a novel set of microsatellite markers for the endangered undulate ray (*Raja undulata*) and subsequently assess their power and informativeness for *ex situ* conservation of this species. Genomic DNA, extracted from a tissue sample, was successfully used to generate a sequencing library, and bioinformatics and laboratory techniques were employed to discover and optimise seven microsatellite markers from the resulting next-generation sequencing (NGS) data set. In order to undertake genetic analyses of this nature, a reliable source of DNA is required, but often this can come at the cost of distress or harm to the subject. Therefore, non-invasive genetic sampling methods are preferable to invasive tissue, blood or biopsy sampling, particularly for threatened species. Although an initial tissue sample was used for the development of the markers, a minimally-invasive sampling method for the collection of the remaining samples from the captive animals (Lieber et al., 2013) was tested. This technique takes advantage of the mucus secreted by the skin of many elasmobranchs and this study demonstrates the successful isolation of high quality, amplifiable DNA from captive animals. The new markers were used to genotype a small captive population of 35 animals, across 10 UK aquaria, demonstrating that the minimally-invasive sampling methodology was suitable for a study of this nature. Several quality-checking procedures were applied to the markers themselves, such as tests for linkage disequilibrium (LD) or deviations from Hardy–Weinberg Equilibrium (HWE). Evidence of both LD and deviation from HWE was observed in some markers. The deviation from expected HWE can be attributed to the fact that the test population breaks many of the underlying assumptions of HWE, mainly that one should consider a large, unrelated population, which is not the case here. Several statistical analyses of the data were performed, making routine measurements of heterozygosity of the population at these loci, calculating inbreeding coefficients and genetic distance, for example. The

results show rates of heterozygosity at each marker ranging from 0.54–0.94 (average 0.81), implying that when all markers are taken into account, the rate of genetic variation in the captive population is not likely to be significantly lower than the wild population from which it was founded. For comparison, in a similar study (Chapman et al., 2011), seven microsatellite markers were used to measure heterozygosity in an elasmobranch population consisting of 104 individuals of the critically-endangered smalltooth sawfish (*Pristis pectinata*) and discovered an average rate of heterozygosity of 0.83. Heterozygosity rates in wild-caught animals and captive-bred, F1 generation individuals did not show any significant difference, demonstrating that a high proportion of genetic variation has been carried into this generation. Data reporting the proportion of wild-caught individuals that successfully contributed to the F1 generation are unfortunately not available. These measures should be repeated at each new generation and can be interpreted as a proxy for the measure of total variation in the group. The captive-bred *R. undulata* of the present study had an average rate of heterozygosity of 0.83. It is important to note, however, that these results on the captive-bred population only take into account the F1 generation and that any decrease in the rate of heterozygosity will likely become apparent over subsequent generations (Willoughby et al., 2017). Continued monitoring via the methods explained in this study will be critical to continue to evaluate the genetic diversity of the population and to continue to monitor for inbreeding depression. Several aquaria housing private alleles within their cohort have been identified, and this information may be useful for maintaining genetic variation when the breeding plan is developed. While it is common to calculate the likely pedigree (i.e. relatedness) from this type of genetic data, the power to correctly assign offspring to parents will be very low for captive populations with a limited captive population size. In these cases, it is far more informative to directly examine the genetic similarity of individuals. The calculation of Provesti's genetic distance (Prevosti et al., 1975) enabled the visualisation of a proxy measure of dissimilarity between individuals (Figure 2.2) through calculating the absolute genetic distance between each pair of individuals. Figure 2.2 shows no clustering around a particular aquarium or between the wild-caught or captive-bred groupings, indicating the lack of an aquarium effect or differentiation of the F1 generation from the wild individuals. Rather, the individual genotypes suggest a homogeneous mixture with no apparent groupings, or sub-structuring emerging. These results fall within expectations as 50% of the total individuals were wild-caught (17 of 35) and so can be expected to be reasonably

unrelated to one another as they originate from a wild population. Progeny from relatively high admixture would be expected to maintain high levels of variation in the F1 generation and similarly be relatively unrelated to one another (with the exception of siblings, parents-progeny, etc.). This study leads to the recommendation that similar analyses be performed as new individuals are caught, born or moved between aquaria to enable population managers to intervene should a particular group of individuals appear to become distinct from other groups, or when one of the measures, or proxy measures, of variation among individuals begins to fall. With a greater number of microsatellite markers, the work could be extended to include relatedness estimates of a much higher confidence and this would also lead to the production of accurate pedigrees—very useful tools to the community managing these animals, but beyond the scope of this piece of work.

2.8 Conclusion

Ex-situ conservation is a very important management tool and is likely to be increasingly used as the rate of anthropogenic influenced species declines continues to climb (Ceballos et al., 2015). Captive populations must be carefully and strategically managed in order to successfully provide individuals for reintroduction, maintain genetic variation and reduce the negative effects of inbreeding (Frankham et al., 2004). (Janse et al., 2017) succinctly summarised the contemporary elasmobranch populations in European aquaria and identified the requirement for good programme management. This study demonstrates that researchers can move relatively quickly from collecting tissue/swab samples, through designing a novel marker panel to producing quantifiable, genetic data and drawing conclusions regarding the structure of a captive population (the majority of the work on this analysis was performed in a matter of a few months). In the absence of a good quality pedigree or studbook, these techniques should form the minimum requirement when working with *ex situ* populations, and as NGS technologies continue to improve, the number and nature of available markers will also increase, leading to significant gains in the quality of the data available. The power of this particular study was limited by a lack of markers, thus preventing some analyses from being performed. However, from the data generated here, it is evident that the population of undulate rays in UK aquaria do not currently appear to be suffering from any malady resulting from their small population size, and the findings appear to fall in

line with other managed groups of elasmobranchs. The results, however, constitute a time bound observation and are therefore only representative of the population at the time the samples were taken. In conclusion, the study has shown that it is feasible and useful to design and optimise a panel of markers for a small, *ex situ* population and that even with a small number of markers, the resulting data can be informative and help with the management of the population. With these markers available to the community, it is hoped that a better understanding of the captive population in UK aquaria in relation to individuals in European aquaria and in wild populations can be reached. This study forms the basis for further scope of greater scope, encompassing a greater sample size, more sampling sites (aquaria) and more microsatellite markers to increase the statistical power of the analyses.

2.9 Acknowledgements

This research was funded by the University of Manchester Faculty of Life Sciences (FLS), a FLS Business Development Small Award, and the Sea Life Trust. Our thanks go to the DNA Sequencing Facility and the Genomic Technologies Core Facility, both at the University of Manchester (UK), for their expert advice and services and two anonymous reviewers for their helpful comments, which improved the manuscript.

2.10 Conflicts of Interest

The authors declare that they have no conflict of interest.

References

Adamack, A. and Gruber, B. (2014). Popgenreport: simplifying basic population genetic analyses in R. *Methods in Ecology and Evolution*, 5(4):384–387.

Appukuttan, K. and Nair, K. (1988). Shark resources of India, with notes on biology of a few species. In Joseph, M., editor, *The First Indian Fisheries Forum, Proceedings*, pages 173–184. Asian Fisheries Society, Indian Branch, Mangalore, India.

Baum, J., Myers, R., Kehler, D., Worm, B., Harley, S., and Doherty, P. (2003). Collapse and conservation of shark populations in the northwest Atlantic. *Science*, 299(5605):389–392.

Blacket, M., Robin, C., Good, R., Lee, S., and Miller, A. (2012). Universal primers for fluorescent labelling of PCR fragments–an efficient and cost effective approach to genotyping by fluorescence. *Molecular Ecology Resources*, 12(3):456–63.

Bolger, A., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–20.

Botsford, L., Castilla, J., and Peterson, C. (1997). The management of fisheries and marine ecosystems. *Science*, 277(5325):509–515.

Castoe, T., Poole, A., de Koning, A., Jones, K., Tomback, D., Oyler-McCance, S., Fike, J., Lance, S., Streicher, J., Smith, E., and Pollock, D. (2015). Correction: Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PLOS ONE*, 7(2):e30953.

Ceballos, G., Ehrlich, P., Barnosky, A., García, A., Pringle, R., and Palmer, T. (2015). Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Science Advances*, 1(5):e1400253.

Chapman, D., Simpfendorfer, C., Wiley, T., Poulakis, G., Curtis, C., Tringali, M., Carlson, J., and Feldheim, K. (2011). Genetic diversity despite population collapse in a critically endangered marine fish: The smalltooth sawfish (*Pristis pectinata*). *Journal of Heredity*, 102(6):643–52.

Christie, M., Marine, M., French, R., and Blouin, M. (2012). Genetic adaptation to captivity can occur in a single generation. *Proceedings of the Natural Academy of Sciences of the United States of America*, 109(1):238–242.

Coelho, R., Bertozzi, M., Ungaro, N., and Ellis, J. (2009). *Raja undulata.* the IUCN red list of threatened species 2009: e.t161425a5420694.

Compagno, L. (1990). Alternative life-history of cartilaginous fishes in time and space. *Environmental Biology of Fishes*, 28(1-4):33–75.

Dawson, T., Jackson, S., House, J., Prentice, I., and Mace, G. (2011). Beyond predictions: Biodiversity conservation in a changing climate. *Science*, 332(6025):52–58.

Dunn, S., Barnowe-Meyer, K., Gebhardt, K., Balkenhol, N., Waits, L., and Byers, J. (2010). Ten polymorphic microsatellite markers for pronghorn (*Antilocapra americana).* *Conservation Genetics Resources*, 2(1):81–84.

Earnhardt, J. (2010). The role of captive populations in reintroduction programs. In Kleiman D.G., Thompson K.V., B. C., editor, *Wild mammals in captivity: principles and techniques for zoo management.* University of Chicago Press, Chicago, IL.

Ellis, J., McCully, S., and Brown, M. (2012). An overview of the biology and status of undulate ray (*Raja undulata*) in the north-east Atlantic Ocean. *Journal of Fish Biology*, 80(5):1057–74.

Fernández, J., Toro, M., and Caballero, A. (2004). Managing individuals' contributions to maximise allelic diversity maintained in small, conserved populations. *Conservation Biology*, 18(5):1358–1367.

Ferretti, F., Myers, R., Serena, F., and Lotze, H. (2008). Loss of large predatory sharks from the Mediterranean Sea. *Conservation Biology*, 22(4):952–64.

Ferretti, F., Worm, B., Britten, G., Heithaus, M., and Lotze, H. (2010). Patterns and ecosystem consequences of shark declines in the oceans. *Ecology Letters*, 13(8):1055–71.

Frankham, R., Ballou, J., and Briscoe, D. (2004). *A Primer of Conservation Genetics.* Cambridge University Press, Cambridge, USA.

Frisk, M., Miller, T., and Fogarty, M. (2001). Estimation and analysis of biological parameters in elasmobranch fishes: a comparative life history study. *Canadian Journal of Fisheries and Aquatic Sciences*, 58(5):969–981.

Gibson, C., Valenti, S., Fowler, S., and Fordham, S. (2006). The conservation status of northeast Atlantic Chondrichthyans; report of the IUCN shark specialist group northeast Atlantic regional red list workshop. *VIII + 76pp. IUCN SSC Shark Specialist Group.*

Griffiths, S., Fox, G., Briggs, P., Donaldson, I., Hood, S., Richardson, P., Leaver, G., Truelove, N., and Preziosi, R. (2016). A Galaxy-based bioinformatics pipeline for optimised, streamlined microsatellite development from Illumina next-generation sequencing data. *Conservation Genetics Resources*, 8(4):481–486.

Janse, M., Zimmerman, B., Geerlings, L., Brown, C., and Nagelkerke, L. (2017). Sustainable species management of the elasmobranch populations within European aquariums: a conservation challenge. *Journal of Zoo and Aquarium Research*, 5(4).

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11):1403–5.

Jombart, T. and Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome wide SNP data. *Bioinformatics*, 27(21):3070–1.

Koressaar, T. and Remm, M. (2007). Enhancements and modifications of primer design program Primer3. *Bioinformatics*, 23(10):1289–91.

Kruskal, J. (1964). Multidimensional scaling by optimising goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.

Lacy, R. (2012). Achieving true sustainability of zoo populations. *Zoo Biology*, 32(1):19–26.

Lieber, L., Berrow, S., Johnston, E., Hall, G., Hall, J., Gubili, G., Sims, D., Jones, C., and Noble, L. (2013). Mucus: aiding elasmobranch conservation through non-invasive genetic sampling. *Endangered Species Ressearch*, 21:215–222.

Nakano, H. (1999). Fishery management of sharks in Japan. In Shotton, R., editor, *Case studies of the management of elasmobranch fisheries.* Fisheries and Aquaculture Department, Food and Agriculture Organization of the United Nations.

Oksanen, J., Guillaume Blanchet, F., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P., O'Hara, R., Simpson, G., Solymos, P., Henry, M., Stevens, H., Szoecs, E., and Wagner, H. (2017). vegan: Community ecology package. R package version 2.4-4.

Oliver, S., Braccini, M., Newman, S., and Harvey, E. (2015). Global patterns in the bycatch of sharks and rays. *Marine Policy*, 54:86–97.

Pelletier, F., Reale, D., Watters, J., Boakes, E., and Garant, D. (2009). Value of captive populations for quantitative genetics research. *Trends in Ecology and Evolution*, 24(5):263–270.

Pew, J., Muir, P., Wang, J., and Frasier, T. (2015). related: an R package for analysing pairwise relatedness from codominant molecular markers. *Molecular Ecology Resources*, 15(3):557–61.

Pimm, S., Jenkins, C., Abell, R., Brooks, T., Gittleman, J., Joppa, L., Raven, P., Roberts, C., and Sexton, J. (2014). The biodiversity of species and their rates of extinction, distribution and protection. *Science*, 344(6187).

Prevosti, A., Ocaña, J., and Alonso, G. (1975). Distances between populations of *Drosophila subobscura*, based on chromosome arrangement frequencies. *Theoretical and Applied Genetics*, 45(6):231–241.

Price, M. and Hadfield, M. (2013). Population genetics and the effects of a severe bottleneck in an *ex situ* population of critically endangered Hawaiian tree snails. *PLOS ONE*, 9(12):e114377.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raymond, M. and Rousset, F. (1995). GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, 86(3):248–249.

Rogers, J. (1972). Measures of genetic similarity and genetic distance. In Wheeler, M., editor, *Studies in Genetics VII*. The University of Texas, Austin, Texas.

Rose, D. (1998). Shark fisheries and trade in the Americas, volume 1: North America. World Wildlife Fund(TRAFFIC Report).

Rousset, F. (2008). genepop'007: a complete reimplementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, 8(1):103–6.

Shepherd, T. and Myers, R. (2005). Direct and indirect fishery effects on small coastal elasmobranchs in the northern Gulf of Mexico. *Ecology Letters*, 8(10):1095–1104.

Smith, S., Au, D., and Show, C. (1998). Intrinsic rebound potentials of 26 species of Pacific sharks. *Marine Freshwater Research*, 49(7):663–678.

Stevens, J., Bonfil, R., Dulvy, N., and Walker, P. (2000). The effects of fishing on sharks, rays and chimaeras (chondrichthyans), and the implications for marine ecosystems. *ICES Journal of Marine Science*, 57(3):476–494.

Untergasser, A., Cutcutachem, I., Koressaarm, T., Ye, J., Faircloth, B., Remm, M., and Rozen, S. (2012). Primer3 - new capabilities and interfaces. *Nucleic Acids Research*, 40(15):e115.

Walker, P. and Hislop, J. (1998). Sensitive skates or resilient rays? Spatial and temporal shifts in ray species composition in the central and northwestern North Sea between 1930 and the present day. *ICES Journal of Marine Science*, 55(3):392–402.

Willoughby, J., Fernandez, N., Lamb, M., Ivy, J., Lacy, R., and DeWoody, J. (2014). The impacts of inbreeding, drift and selection on genetic diversity in captive breeding populations. *Molecular Ecology*, 24(1):98–110.

Willoughby, J., Ivy, J., Lacy, R., Doyle, J., and DeWoody, J. (2017). Inbreeding and selection shape genomic diversity in captive populations: Implications for the conservation of endangered species. *PLOS ONE*, 12(4):e0175996.

Witzenberger, K. and Hochkirch, A. (2011). *Ex situ* conservation genetics: a review of molecular studies on the genetic consequences of captive breeding programmes for endangered animal species. *Biodiversity and Conservation*, 20(9):1843–1861.

2.11 Figures and Tables

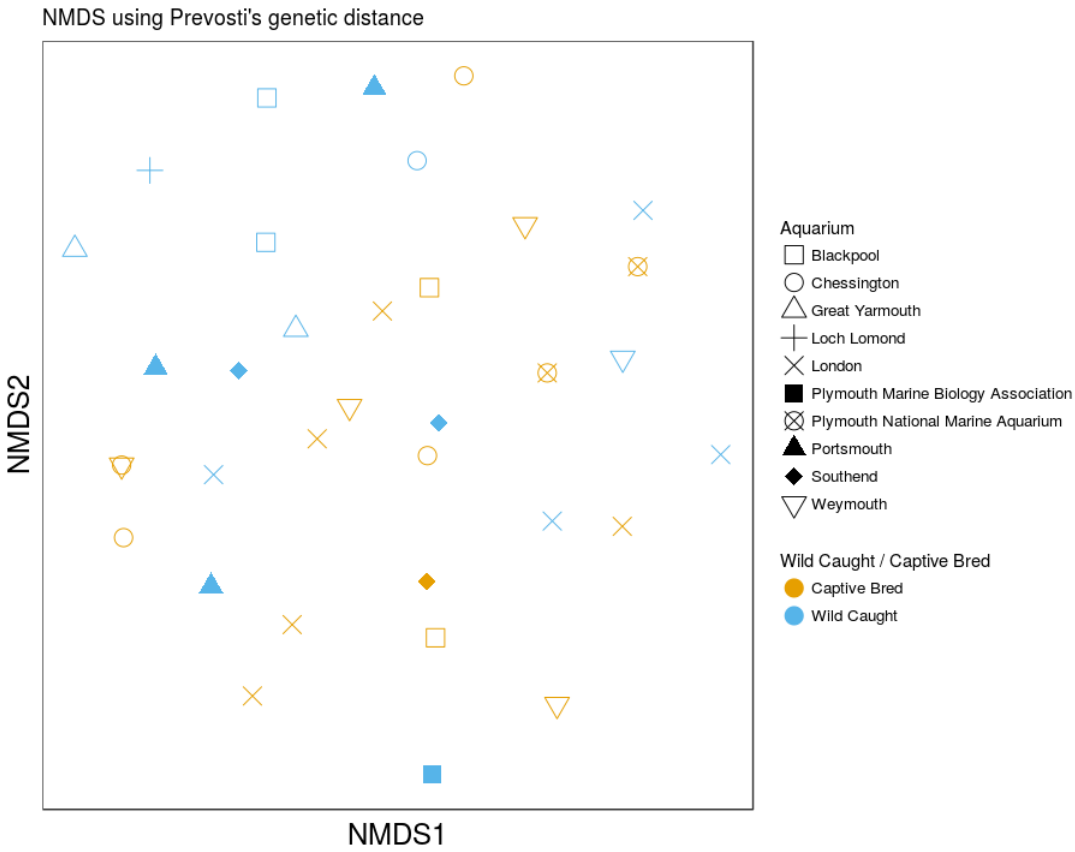


Figure 2.2

Table 2.1

Aquarium	Sample Number	Wild-Caught	Captive-Bred	Private Alleles
Sea Life London Aquarium	9	4	5	2
Weymouth Sea Life Adventure Park	5	1	4	2
Sea Life Blackpool	4	2	2	0
Sea Life Chessington	5	1	4	0
Sea Life Adventure, Southend	3	2	1	1
Sea Life Great Yarmouth	2	2	0	2
Sea Life Loch Lomond	1	1	0	0
Blue Reef Aquarium, Portsmouth	3	3	0	1
National Marine Aquarium, Plymouth	2	0	2	3
Marine Biology Association, Plymouth	1	1	0	0

Table 2.2

Locus ID, nucleotide motif, number of alleles (NA), size range of fragments (SR), PCR annealing temperature (T_A), expected (H_{EXP}) and observed (H_{OBS}) heterozygosity, P-value from testing for Hardy–Weinberg equilibrium (P_{HWE}), number of individuals tested (N), primer nucleotide sequences ($5'$ to $3'$ orientation) and raw sequence accession numbers. *Marker RU13 not used in this study due to deviation from expected HWE values.

Table 2.2

Locus	Motif	NA	SR (bp)	T_A	H_{EXP}	H_{OBS}	P_{HWE}	N	Primers (5′- 3′)	Accession
Ru_02	AAGAGG	10	347-419	60	0.808	0.800	0.0180	35	CCCTGTTCTCCTGCTCTCCATTACC	MH049873
									CTCTCCCTATAGCTCAGGCCTTCGG	
Ru_03	ACTGCC	10	412-463	60	0.827	0.882	0.0694	34	CATTCACAACTGCAGTCCAATGTCC	SRP134840
									TCTGCTGTCAAGCTGTTGTGTCAGG	
Ru_08	AGGTG	13	351-415	60	0.887	0.800	0.0113	35	TGAGGAATTCATTGCCACAAACTGC	MH049874
									TCCTCTCACATAACCCTGTGTATGCC	
Ru_09	ATAG	22	209-385	60	0.945	0.939	0.1463	33	TCTTTGCTCCTACCGGTTCTTCTCG	MH049875
									CAGAACAAGGCTTGGTGGTCTTGG	
Ru_13*	ACAG	9	317-373	60	0.787	0.313	0	32	CATTCTTAACAGGGCAGCTACTTGTGG	MH049876
									AAAGATTGGTAGGAAGATGGATCGG	
Ru_14	AGGC	8	277-313	60	0.754	0.882	0.7937	34	ACCTCGAAACCGCCATTAAGAATCC	MH049877
									CTGCATGTTATCGAGCAATCAGTCG	
Ru_20	ACAG	9	374-407	60	0.846	0.886	0.1317	35	GGACACTTGACACAGCTTTGGTCTCC	MH049878
									GGGAGTTACCTTCATGGTGAGACAGG	
Ru_21	AAT	5	373-388	60	0.682	0.543	0.1631	35	CATGACTGGGGCTAGAAGGTGTTGC	MH049879
									GTTAGAGCAGTCCGCCATGAAGGG	

Table 2.3

Species	Locus Name							
	Ru02	Ru03	Ru08	Ru09	Ru13	Ru14	Ru20	Ru21
Raja microcellata	341-419	412-463	351-432	209-385	317-373	277-376	374-407	373-388
Raja brachyura	347-377	408-463	351-428	204-385	317-419	277-391	374-407	373-388
Raja montagui	347-364	412-483	351-415	209-385	317-373	277-313	374-422	373-388
Raja clavata	343-419	412-463	351-415	209-385	285-373	277-313	374-407	373-388

Chapter 3

Microsatellite Marker Design Methods Using Next-Generation Sequence Data.

3.1 A Galaxy-based bioinformatics pipeline for optimised, streamlined microsatellite development from Illumina next-generation sequencing data.

3.1.1 Brief Note

With my colleagues we developed and optimised a web-based tool which allows users around the world interested in developing microsatellite marker panels to implement a workflow optimised in our laboratory. The tool is available online: `https://palfinder.ls.manchester.ac.uk/`

My contribution to this project was to develop a quality control method which is implemented in the tool (PANDAseq_QC, detailed below) and also to provide the automation of the entire workflow via Python scripting. My scripts form the backbone of the online tool, accessed via "wrapper" scripts to which I also contributed which are used by the Galaxy environment. The tool is now available in the Galaxy Toolshed for download.

3.1.2 Publication Reference

This piece of work is published at: Griffiths, S.M., Fox, G., Briggs, P.J., Donaldson, I.J., Hood, S., Richardson, P., Leaver, G.W., Truelove, N.K. and Preziosi, R.F. (2016) A Galaxy-based bioinformatics pipeline for optimised, streamlined microsatellite development from Illumina next-generation sequencing data. *Conservation Genetics Resources*. 8(4): 481-486.

Conservation Genet Resour (2016) 8:481–486
DOI 10.1007/s12686-016-0570-7

METHODS AND RESOURCES ARTICLE

A Galaxy-based bioinformatics pipeline for optimised, streamlined microsatellite development from Illumina next-generation sequencing data

Sarah M. Griffiths[1] · Graeme Fox[1] · Peter J. Briggs[2] · Ian J. Donaldson[2] · Simon Hood[3] · Pen Richardson[3] · George W. Leaver[3] · Nathan K. Truelove[4] · Richard F. Preziosi[1]

Received: 19 February 2016 / Accepted: 1 July 2016 / Published online: 2 August 2016

Figure 3.1

3.2 Multi-individual Microsatellite identification: a multiple genome approach to microsatellite design (MiMi).

Graeme Fox,[1] Richard F. Preziosi,[1] Rachael A. Antwis,[2] Fraser J. Combe,[3] W. Edwin Harris,[4] Ian R. Hartley,[5] Andrew C. Kitchener,[6] Selvino R. de Kort,[1] Anne-Isola Nekaris,[7] Milena Benavides-Serrato,[1,8] Jennifer K. Rowntree,[1]

[1]Ecology and Environment Research Centre, Department of Natural Sciences, Manchester Metropolitan University, Chester Street, Manchester, United Kingdom, M1 5GD

[2]School of Environment and Life Sciences, University of Salford, Salford, M5 4WT, UK.

[3]Kansas State University, Division of Biology, Manhattan, KS, USA

[4]Crop & Environment Sciences, Harper Adams University, Shropshire, TF10 8NB, UK.

[5]Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ

[6]Department of Natural Sciences, National Museums Scotland, Chambers Street, Edinburgh, EH1 1JF, UK.

[7]Department of Social Sciences, Faculty of Humanities and Social Sciences, Oxford Brookes University, Oxford, OX3 0BP, UK.

[8]Universidad Nacional de Colombia, Sede Caribe-CECIMAR Calle 25 2-55, Playa Salguero, Colombia.

Key words: Microsatellite design, High-throughput sequencing, Short Tandem Repeat (STR), *in silico* quality control, Polymorphic loci detection, Cost-effective marker development.

Author Contributions: GF, RFP and JKR conceived and designed the project; FJC, WEH, IRH, ACK, SRdK, AIN, MBS collected and/or provided the samples; GF, FJC and MBS performed the lab work; RAS assisted with the sequencing of some samples; GF performed the computer programming; GF performed the data analysis; GF wrote the chapter.

3.3 Publication Reference

This chapter is published at:

Fox, G., Preziosi, R.F., Antwis, R.E., Benavides-Serrato, M., Combe, F.J., Harris, W.E., Hartley, I.R., de Kort, S.R., Nekaris, A. and Rowntree, J.K. (2019) *Molecular Ecology Resources* DOI: https://doi.org/10.1111/1755-0998.13065

Received: 6 July 2018 | Revised: 17 June 2019 | Accepted: 18 June 2019

DOI: 10.1111/1755-0998.13065

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Multi-individual microsatellite identification: A multiple genome approach to microsatellite design (MiMi)

Graeme Fox[1] | Richard F. Preziosi[1] | Rachael E. Antwis[2] | Milena Benavides-Serrato[1,3] | Fraser J. Combe[1,4] | W. Edwin Harris[1,5] | Ian R. Hartley[6] | Andrew C. Kitchener[7] | Selvino R. de Kort[1] | Anne-Isola Nekaris[8] | Jennifer K. Rowntree[1]

Figure 3.2

3.4 Abstract

Bespoke microsatellite marker panels are increasingly affordable and tractable to researchers and conservationists. The rate of microsatellite discovery is very high within a shotgun genomic data set, however extensive laboratory testing of markers is required for confirmation of amplification and polymorphism. By incorporating shotgun next-generation sequencing data sets from multiple individuals of the same species, we have developed a new method for the optimal design of microsatellite markers. This new tool allows us to increase the rate at which suitable candidate markers are selected by 58% in direct comparisons and facilitate an estimated 16% reduction in costs associated with producing a novel microsatellite panel. Our method enables the visualisation of each microsatellite locus in a multiple sequence alignment allowing several important quality checks to be made. Polymorphic loci can be identified and prioritised. Loci containing fragment length altering mutations in the flanking regions, which may invalidate assumptions regarding the model of evolution underlying variation at the microsatellite, can be avoided. Priming regions containing point mutations can be detected and avoided, helping to reduce sample site marker specificity arising from genetic isolation, and the likelihood of null alleles occurring. We demonstrate the utility of this new approach in two species: an echinoderm and a bird. Our method makes a valuable contribution towards minimising genotyping errors and reducing costs associated with developing a novel marker panel. The Python script to perform our method of Multi-individual Microsatellite identification (MiMi) is freely available from GitHub (https://github.com/graemefox/mimi).

3.5 Introduction

Microsatellites, short tandem repeats (STRs) or short simple repeats (SSRs), are exceptionally polymorphic repetitive regions of DNA found throughout the genomes of both eukaryotic and prokaryotic species (Bhargava and Fuentes, 2010; Rose and Falush, 1998). High rates of polymorphism, along with co-dominance and Mendelian inheritance, make them ideal markers for use in studies of population genetics (Abdul-Muneer, 2014; Goldstein and Pollock, 1997). Microsatellites have been the most popular choice of genetic marker for several decades in ecology, conservation and evolutionary research, and are extensively used in contemporary studies of population genetics, parentage and

kinship identification, evolutionary processes and genetic mapping (Vieira et al., 2016; Ribout et al., 2019). Although single nucleotide polymorphism (SNP) markers have become increasingly popular markers for population genetics, microsatellites remain a common choice due to well-documented methodologies, ease of application, low equipment demands and well-developed statistical analyses. Furthermore, there remain scenarios where SNPs are not practical for use, or microsatellites are preferred (Zhan et al., 2016). For example, the management of captive populations has benefited enormously by the inclusion of genetic information (Fox et al., 2018; Witzenberger and Hochkirch, 2011), which must be continually updated as small numbers of new individuals are added to collections or produced through mating. In these cases, it is impractical to perform repeated SNP analysis on small numbers of samples due to the expense associated with next-generation sequencing (NGS) to acquire high coverage SNPs. Conversely, once a microsatellite panel has been developed, additional individuals can be genotyped using the existing markers very quickly, and at very low cost (Puckett, 2016). Where non-invasive sampling methods are required, for example because a species is of conservation concern (e.g. (Fox et al., 2018)), it may prove to be impossible to acquire sufficient high molecular weight DNA to perform NGS for SNP genotyping. In contrast, microsatellite analysis is forgiving of low DNA template input, and many contaminants that may disrupt NGS library preparation can simply be diluted out prior to amplification. A simple literature search in Google Scholar indicated the publication of approximately 2000 new microsatellite marker panels in 2018, suggesting that microsatellites are still very popular genetic markers, and we predict they will continue to be used extensively in conservation and ecology well into the future.

Ecological and conservation studies are often focused upon non-model species for which genetic markers are not available. The combination of affordable NGS and freely available bioinformatics tools can be used to identify tens of thousands of potential markers in a matter of days. Where probes were once used to target repeat regions of genetic code (Bloor et al., 2001), shotgun genome sequencing does not require any prior knowledge of the genome, and is considered a non-targeted approach (Davey et al., 2011). Instead, random fragments of genomic DNA are sequenced, a fraction of which include SSRs within the length of the sequencing read. Free, open source software packages are available to detect SSRs and design suitable PCR primers to amplify the appropriate region of the genome; often referred to as the "seq-to-SSR" approach (Castoe et al., 2015; Griffiths et al., 2016). These developments and the increasing availability of NGS technology globally, brings

microsatellite marker discovery within the reach of ever more research laboratories as the cost-per-base of NGS continues to decrease (Koboldt et al., 2013), even for applied, species-focused conservation research with limited funding. Thus, the development of bespoke microsatellite marker panels has become commonplace.

The use of microsatellite markers is reliant upon variation in PCR product fragment length, and therefore microsatellites must be amplifiable by PCR, and must contain fragment length altering polymorphisms within the repetitive stretch of SSR sequence. Despite improvements delivered by NGS, the optimisation of a bespoke microsatellite panel remains a time consuming and costly process, largely because the primer pair for each potential marker still requires manual laboratory confirmation of both successful amplification and the presence of multiple alleles at the locus (Bloor et al., 2001). Typically, the development of a microsatellite marker is performed through the discovery of a microsatellite locus in a single individual, followed by analysis of the locus in several more individuals to test for consistent amplification and variation in PCR fragment size (Abdelkrim et al., 2009). The main contributors to the cost of developing a panel of microsatellite markers are the NGS reagents, PCR reagents, PCR oligos, capillary electrophoresis, size standards and staff time. Improvements that enable reductions in cost or time associated with marker development will contribute to microsatellite markers becoming more widely available to ecological and conservation researchers.

Here we present a new conceptual approach to microsatellite marker design, demonstrated with a new bioinformatics technique in application to seq-to-SSR workflows. This technique is designed to improve the rate at which loci that are identified can be successfully amplified by PCR and produce informative genotype data. The innovation in our approach is the incorporation of information from the genomes of multiple individuals. This allows the in silico detection of polymorphic loci and the detection of several other important characteristics of a putative microsatellite marker, only detectable through multiple genome analysis. We demonstrate that this method reduces the number of markers that must be tested for polymorphism in the laboratory, and achieves an improved rate of successful marker development. Furthermore, our methods also minimise factors known to increase allelic dropout and invalidate genotyping results based upon molecular weight of PCR fragments. We refer to this technique as Multi-individual Microsatellite identification (MiMi). Here, we develop microsatellite markers using MiMi in two species: the green sea urchin (*Psammechinus miliaris*) and the Eurasian blue tit (*Cyanistes caeruleus*). For comparison, we also

present the success rates of microsatellite development in *P. miliaris* and *C. caeruleus*, and in two other species (*Tragelaphus eurycerus isaaci* and *Nycticebus pygmaeus*), which were designed using a traditional microsatellite design method (Castoe et al., 2015; Griffiths et al., 2016). The results from the successful development of each panel of markers, combined with our refined bioinformatics method, provide a strong case for the utility of the MiMi concept and the value to microsatellite marker development.

3.6 Materials and Methods

3.6.1 DNA Extraction and Sequencing

Prior to DNA extraction, all samples (Appendix 3: Table S1) were stored in 100% ethanol and stored at 4°C. Genomic DNA was extracted from samples using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) or the E.Z.N.A. Mollusc DNA Kit (Omega bio-tek, Georgia, USA) (Appendix 3: Table S2). High quality and high molecular weight genomic DNA (determined by gel electrophoresis) was diluted to 2.5ng/µL and sequenced on an Illumina MiSeq (Illumina, San Diego, USA), using the Illumina Nextera XT library preparation reagents (Illumina, San Diego, USA). Paired-end, shotgun genomic DNA sequencing was performed using the Illumina MiSeq Reagent Kit v2/v3. MiMi analysis was conducted on eight individuals of each species (*P. miliaris* and *C. caeruleus*) which were indexed, pooled and sequenced on a flowcell, per species. For traditional microsatellite detection, single samples of each species (*T. eurycerus isaaci* and *N. pygmaeus*) were individually indexed, pooled and sequenced along with other species not used in this study (Appendix 3: Table S2). Both methods were not tested for all species, due to these microsatellite markers being designed for active research projects that progressed beyond marker development as the MiMi method was being developed and iterated upon.

3.6.2 MiMi Microsatellite Detection Methodology

Microsatellite markers were initially designed in data from each sample using the published pal_finder (Castoe et al., 2015; Griffiths et al., 2016); a traditional design method using the data of a single individual. A novel quality control procedure was developed for those data sets in which multiple individuals of the same species were sequenced (two species) with the aim of identifying polymorphic loci, filtering out primer

pairs containing point mutations within the priming regions, and avoiding other potential issues with a locus including non-specific primer binding and insertion/deletion mutations in the flanking regions. Eight individuals per species were sequenced and the data pertaining to each individual first passed separately through the traditional design method. The eight individual output files then become the input for the novel method: Multi-individual Microsatellite identification (MiMi). MiMi takes the primer sequences developed in each individual and checks for their presence in the data of every other individual. Primer pairs for which the forward primer appeared in more than 33% of the individuals were selected and all reads containing the exact primer sequence compiled into an MSA file with the FASTA format. The MSA files were aligned using the MUSCLE alignment algorithm (Edgar, 2004) and putative loci automatically filtered to remove monomorphic loci, low quality 'gapped' alignments and loci containing sequence mutations within the primer binding sites. Loci passing all filters are retained as high quality loci and loci passing some filters but lacking enough information to confidently pass all filters are retained as good quality loci. Both high quality and good quality loci are each ranked by the size range in alleles detected. A log file is produced detailing loci which have been removed by each filter. A Python script implementing the MiMi tool is available to download and run from https://github.com/graemefox/mimi.

3.6.3 Optimisation of Potential Markers

Primer pairs developed under either design method were tested in 5µL reactions using the Type-it Microsatellite PCR Kit (Qiagen, Hilden, Germany) using the standard protocol and thermal cycling parameters (5 mins at 95°C, 25-28*30s at 95°C, 90s at 60°, 30s at 72°, 30 mins at 60°). Only a single annealing temperature (60°C) was tested, as Primer3 (Koressaar and Remm, 2007; Untergasser et al., 2012) which is used during the published marker design process (Griffiths et al., 2016), had been configured specifically for these PCR reagents and a primary goal of this method was to avoid time consuming annealing temperature optimisation. A marker was given successful amplification status if clean PCR products were clearly visible on a 2% agarose gel in the 100-1000bp range for six or more individuals out of eight tested. Fluorescent dyes (6-FAM, TAMRA, HEX, PET) were added to PCR products using a universal tail technique (Blacket et al., 2012). Fragment length was determined using an ABI 3730 DNA Analyzer capillary sequencer with GeneScan 500 LIZ dye Size Standard (ThermoFisher and analysed using Genemapper 5.0 software (all ThermoFisher Scientific, Carlsbad, USA). We define an

informative marker as one that produces clearly interpretable electropherogram traces after capillary electrophoresis and is polymorphic in terms of PCR fragment length between multiple individuals.

3.7 Results

Of the markers which passed each set of quality controls, we were able to optimise amplifiable and informative markers at a rate of 47.9% using the traditional design method, and 86.6% using MiMi. Comparisons between average rates of successful amplification and production of informative loci for each marker design method demonstrated a marked increase in both measures when MiMi was applied. In *P. miliaris* and *C. caeruleus*, markers were designed using both the traditional methodology and the MiMi methodology. A direct comparison between these two methods shows a very notable increase in both the rate of amplification success and the rate of development of informative markers (Figure 3.3). In two further species (*T. eurycerus isaaci* and *N. pygmaeus*), markers were designed using only the traditional methodology. Rates of success for these species are presented as further evidence of a baseline of microsatellite design against which the MiMi method can be compared (Table 3.1). Unsuitable markers were removed at each filtering stage, reducing hundreds of thousands of possible markers designed by pal_finder, to a fewer than a hundred identified as high- or good-quality using MiMi (Table 3.2). Where MiMi was applied, the number of individuals sharing each common primer sequence ranged from three to seven (Figure 3.4). In the two example MiMi data sets presented here, 5% of potential loci were detected in sufficient individuals to allow further analysis by MiMi.

Automated analysis of MSA files allowed the identification and removal of loci with mutations within the primer binding sites (Appendix 3: Figures S1a and S1b) and loci showing very low alignment quality. Low alignment quality is indicative of a locus potentially containing fragment length altering polymorphisms (insertions/deletions) between the primer binding sites but outside the microsatellite locus itself (Appendix 3: Figure S1c) or non-specific primer binding. Monomorphic loci were also removed (Appendix 3: Figures S1d and S1e). Of the markers which MiMi detected in multiple individuals, we were able to discount 79.3% of potential loci as unsuitable for microsatellite analysis (Table 3.3). High quality loci (those which exclusively showed evidence of positive characteristics) were detected at a rate of 4.5%, and good quality loci (those which did not show any evidence of negative characteristics, but did not have

enough data to confidently pass all filters) were detected at an average rate of 16.1%.

Whilst the full MiMi method requires more data than the traditional approach detailed here (we recommend eight individuals to be sequenced using the capacity of an entire MiSeq flowcell, although fewer samples are possible), the reduction in time spent in the lab, and associated savings, justifies the larger outlay in initial sequencing costs. A recent Illumina MiSeq run cost approximately $2330, and using MiMi we recorded that 90% of the primer pairs chosen to be tested were successfully developed as informative microsatellite markers (Table 3.1, data set #2). Using the traditional method, sequencing costs were significantly less, as only a fraction (12.5%) of the capacity of a MiSeq sequencing flowcell was required, but only 38% of primer pairs tested were ultimately found to be informative markers (Table 3.1, data set #5). The reduction in time and laboratory expense associated with investing in "failed markers" (inconsistent amplification / non-polymorphic loci) ultimately results in a net saving when using MiMi. Based on our estimated rate of successful marker development, a project to develop a panel of 20 optimised markers over a two-week period using the MiMi methodology would cost less than using the traditional methodology over a four week period (16% reduction in total cost, 50% reduction in staff costs only, 19% increase in reagent costs only; Appendix 3: Tables S3 S4). The most significant savings will be in researcher time spent screening loci, which was approximately 50% less using MiMi.

3.7.1 Description of Output Files

The outputs from the MiMi method are two tab separated tables containing details of the loci that have passed the quality control processes, a log file detailing which loci were removed under which quality control conditions, and a per-locus MSA file in the FASTA format. The output tables gives the following information for each locus: forward primer sequence; reverse primer sequence; number of alleles at the locus; number of individuals in which the locus was sequenced in the data set; a description of the alleles found (the repeat motif and the number of repeats) and the predicted size range of amplicons produced using the PCR primers. A file named "MiMi_output_all_loci.txt" gives details of every locus which MiMi was able to detect in multiple individuals (above the user-defined threshold). A file named "MiMi_output_filtered_loci.txt" gives just those loci which were able to pass all quality control filters as either high- or good quality. A log file is created detailing which loci were removed under which quality control conditions. Examples of the "MiMi_output_filtered_loci.txt" files resulting from the the

MiMi analysis of *C. caeruleus* and *P. miliaris* (Table 3.1 data set #1 and data set #2, respectively) is presented in Appendix 3, Tables S5(a) and S5(b) respectively. Three MSA files per locus are created: one containing the raw sequences from the input data that were found to contain the locus within the length of the read (ending .fastq); one containing these reads after alignment by MUSCLE (ending .aln) and one containing aligned reads trimmed to the position of the forward primer (ending .trimmed). The main section of the MSA file name is the forward primer sequence of the locus.

3.8 Discussion

MiMi has proved to be a fast, cost effective approach to identification and characterisation of microsatellite markers using genomic sequence data from multiple individuals. The application of a microsatellite-picking tool such as pal_finder typically results in tens of thousands of potential loci, and therefore it makes logical sense to attempt to apply *in silico* marker optimisation methods over laboratory optimisation, to increase the efficiency in identifying informative loci. MiMi is the first tool, to our knowledge, that allows this range of important characteristics to be observed at the marker design stage (but see (Nichols et al., 2018)). In a direct comparison between the traditional and MiMi methods, we show that the application of MiMi resulted in a 58% increase in the rate of identification of informative microsatellite markers, facilitating a 16% reduction in costs associated with the development of a microsatellite marker panel. To provide a 'baseline' value of microsatellite design success, we also provide success rates for two species which only used the traditional methodology. Although not a true comparison, it appears that MiMi can be expected to produce amplifiable, informative markers at a consistently higher rate than the traditional methodology, facilitating an increase from 57-60% (Table 3.1, data sets #3 and #4) to 80-90% (Table 3.1, data sets #1 and #2). We feel certain that an increase of this order of magnitude, and the associated reduction in costs associated with the testing of markers which ultimately fail, fully justifies the slight increase in sequencing costs associated with MiMi.

The incorporation of multiple genomes and construction of an MSA for each microsatellite locus allows several important quality checks to be made of each locus and facilitates notable increases in both the rate of successful amplification by PCR, and the development of informative markers. Nucleotide polymorphisms and INDEL mutations within the forward or reverse primer binding site can cause issues with inconsistent or

failed PCR amplification, potentially resulting in allelic dropout (Silva et al., 2017), and can also lead to an increase in the frequency of null alleles (Rico et al., 2017). Allelic dropout can present a significant problem during microsatellite analysis, causing decreased estimates of observed heterozygosity and increased estimates of inbreeding in the population (Wang et al., 2012). Two main causes of allelic dropout have been shown; variation at primer binding site (Silva et al., 2017) and PCR product size (particularly problematic for markers with large repeat counts), (Sefc et al., 2003). Through the construction of each MSA we were able to use MiMi to confirm that primer-binding sites show strong sequence conservation, albeit in only a small subset of samples, thus minimising the likelihood that a putative marker would exhibit an elevated rate of allelic dropout caused by mis-priming. Confirmation of sequence conservation in at least one primer-binding site improved the rate at which we were able to amplify loci successfully. If possible, genomes of individuals from a range of putative populations should be included in the MiMi analysis to minimise null allele bias towards a particular sub population (Oosterhout et al., 2005). Analysis of each microsatellite locus in an MSA also allows visualisation of the number of motif repeats, and prioritisation of loci where variation is seen among samples. Rejecting monomorphic loci through MiMi produced an increase in the rate at which we were able to develop informative markers, compared to our own previous experience using other methods, and rates stated in the literature (Zhan et al., 2016). Additionally, MiMi enables one to assess the likelihood of the presence of multiple primer binding sites in the host genome by collating all sequences containing a common primer sequence. Where sequences containing the primer sequence produce low-overlap alignments, it is indicative that the corresponding primer binding site occurs in multiple locations across the genome, and thus that particular primer pair should be avoided to reduce cross-amplification.

Statistical models based upon a particular model of evolution at the microsatellite locus (the stepwise mutation model, for example) rely upon the assumption that the source of variation in fragment size is polymorphism in the number of repeats in the SSR (Dieringer and Schlötterer, 2003). The presence of other fragment length altering mutations between the primer binding sites (excluding the microsatellite itself) is indistinguishable by capillary electrophoresis from 'true' variation at the microsatellite locus (Angers and Bernatchez, 1997; Grimaldi and Crouau-Roy, 1997; Stágel et al., 2009). Markers with fragment length altering mutations outside of the microsatellite locus, potentially invalidate the assumptions of a number of models of microsatellite

evolution, and are therefore avoided in our protocol.

Whilst MiMi does not allow one to state with certainty that a putative marker will not exhibit any of the negative characteristics described (allelic dropout, null alleles arising from population differentiation, non-variable microsatellite loci, cross amplification or invalidation of assumptions of evolutionary model) when comprehensively characterised in a much larger number of samples, the opportunity to identify loci that do exhibit them, and subsequently remove them from analysis, is nevertheless valuable.

Variation in the rate at which loci were removed under each quality control category show the importance of making each check, and that marker development in different taxa may perform differently from one another. In both examples of the application of MiMi here, we were able to remove undesirable loci, that failed at least one quality check. Considering the total markers designed and filtered in both species, we were able to pass many loci (mean: 20.7%) that did not show evidence of these negative characteristics in the eight samples tested.

The success of MiMi is dependent upon the sequence coverage achieved in each sequencing run. Very low sequence coverage would likely result in relatively little overlap in the sequences of each individual, and therefore few loci passing the MiMi filter. The development of a new marker panel is very often performed in non-model species of specialised interest and it is likely that the genome size will be unknown and sequence coverage incalculable (Shikano et al., 2010). MiMi was successfully implemented in the two species tested here (with estimated coverage of 0.57X and 1.20X), suggesting that the method is suitable for genomic data sets with relatively low sequence coverage (Ekblom and Wolf, 2014). The proportion of individuals in which a primer must be detected is user definable, with a minimum of two individuals required for MiMi to provide useful information. Where loci were successfully detected in multiple individuals, we found a negative correlation between the number of potential markers and the frequency at which loci were found in multiple data sets. These frequencies are dependent upon the genome size, and the microsatellite richness of the genome, of the species of interest. Where estimates of genome coverage are approximately 1X or below, removal of duplicate primers/loci from the data set of each individual is recommended (implemented automatically in the published microsatellite design workflow upon which MiMi is based (Griffiths et al., 2016)) as coverage of >1X of a locus in a single individual does not contribute any additional information to the MiMi process. However, where

estimated coverage is significantly >1X, their removal may result in the dismissal of an increased frequency of otherwise useful loci that appear multiple times in the sequence data as a result of the random nature of shotgun sequencing (Bouck et al., 1998). In the event of a low number of markers ultimately being returned, the filter that removes loci appearing more than once in the data can easily be disabled at the web interface of initial design tool (Griffiths et al., 2016). In this case, multiple reads containing the primer sequence from the same biological sample will appear alongside each other in the output MSA, allowing the user to assess the reads as "shotgun duplicates": multiple sequence reads covering the same genomic region of an individual, by chance.

MiMi makes several important assumptions of the characteristics of microsatellite loci investigated in a small number of samples, and infers these are representative of the loci in the wider population. However, this is not always expected to be true (Goldstein et al., 1995) and the removal of otherwise useful markers, under the limiting assumptions of the MiMi quality control process, is likely to happen. For example, SSRs that do not show any variation in number of repeats in the sequence data are removed, but these loci may show variation in the wider population. The ethos behind the MiMi method is to select markers for which we have the most information, rather than seeking to discover as many markers as possible. Given the large numbers of potential markers we derived from the MiMi process, we do not consider the removal of potentially useful markers as a major disadvantage, and these markers can always be added back if needed.

Loci that do show allelic variation are ranked by the range size of the microsatellite repeat number (Goldstein and Schlötterer, 1999), with the assumption that the loci with the largest differences are most likely to be informative markers. A large range in the number of repeats implies that the variation seen at the locus is less likely to be the result of an amplification or sequencing error (Hosseinzadeh-Colagar et al., 2016) but rather be representative of a true, variable microsatellite locus. We conclude that under the assumptions we identify here, the rate and efficiency of informative microsatellite discovery is greatly increased using high-throughput sequencing data in comparison to traditional microsatellite library discovery methods, however the robustness of MiMi should be tested in additional species.

We recommend that eight unrelated individuals are sequenced for MiMi processing for optimal capture of markers exhibiting multiple alleles at microsatellite loci. Whilst it is impossible to state an optimum figure for universal use, due to varying allelic richness in species and populations (Bashalkhanov et al., 2009), in our experience, eight samples

represents an acceptable balance between depth of sequencing coverage and allele rarefaction (Hale et al., 2012). In species where it is not feasible to source eight samples, related or not, due to their extreme scarcity, MiMi is still applicable. MiMi will function beneficially on any number of samples >1, whether related or unrelated. Furthermore, species with extremely large genomes may not perform well due to the limitations of sequencer capacity and the requirement for approximately 1X genome sequence coverage to be achieved. Our method has been tested on Illumina MiSeq data only, but will function on paired-end data, in the .FASTQ format, from any sequencing platform, should additional depth of coverage be required. It is important to note that we are not attempting to detect all, or even most alleles present at a locus. Detecting the presence of multiple alleles (>1) is sufficient to enable MiMi processing. Other influencing factors, such as the sampling of related individuals or populations experiencing low genetic diversity due to historical population bottlenecks, may impact the allelic richness of the samples and therefore the ability of MiMi to detect multiple alleles (Price and Hadfield, 2014).

Methods of genotyping microsatellites by high-throughput sequencing are a promising development and avoid many of the ambiguities inherent in genotyping by capillary electrophoresis (Zhan et al., 2016; Shin et al., 2017). Determination of accurate genotypes by these methods enables many of the additional tests required of a microsatellite marker (tests for linkage disequilibrium, frequency of null alleles, for example) to be carried out using NGS data alone. We envisage that large scale microsatellite studies be performed using two NGS runs: the first using MiMi to discover potentially informative microsatellites; and a second using a high-throughput genotyping method to genotype all experimental samples in one go (De Barba et al., 2016).

3.9 Acknowledgements

With thanks to the Genomic Technologies Core Facility of the University of Manchester for their expertise and services. *P. miliaris* samples were collected by Simon Exley of Queen's University Belfast. Funding for this PhD research comes from Manchester Metropolitan University.

3.10 Conflicts of Interest

The authors declare that they have no conflict of interest.

References

Abdelkrim, J., Robertson, B., Stanton, J., and Gemmell, N. (2009). Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *Biotechniques*, 46(3):185–92.

Abdul-Muneer, P. (2014). Application of microsatellite markers in conservation genetics and fisheries management: recent advances in population structure analysis and conservation strategies. *Genetics Research International*, 2014:691759.

Angers, B. and Bernatchez, L. (1997). Complex evolution of a salmonid microsatellite locus and its consequences in inferring allelic divergence from size information. *Molecular Biology and Evolution*, 14(3):230–8.

Bashalkhanov, S., Pandey, M., and Rajora, O. (2009). A simple method for estimating genetic diversity in large populations from finite sample sizes. *BMC Genetics*, 10(84).

Bhargava, A. and Fuentes, F. (2010). Mutational dynamics of microsatellites. *Molecular Biotechnology*, 44(3):250–66.

Blacket, M., Robin, C., Good, R., Lee, S., and Miller, A. (2012). Universal primers for fluorescent labelling of PCR fragments–an efficient and cost-effective approach to genotyping by fluorescence. *Molecular Ecology Resources*, 12(3):456–63.

Bloor, P., Barker, F., Watts, P., Noyes, H., and Kemp, S. (2001). Microsatellite libraries by enrichment. Online protocol: http://www.genomics.liv.ac.uk/animal/MICROSAT.PDF [Accessed 30/05/2019].

Bouck, J., Miller, W., Gorrell, J., Muzny, D., and Gibbs, R. (1998). Analysis of the quality and utility of random shotgun sequencing at low redundancies. *Genome Research*, 8:1074–1084.

Castoe, T., Poole, A., de Koning, A., Jones, K., Tomback, D., Oyler-McCance, S., Fike, J., Lance, S., Streicher, J., Smith, E., and Pollock, D. (2015). Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PLOS ONE*, 2015(10):e30953.

Combe, F., Taylor-Cox, E., Fox, G., Sandri, T., Davis, N., Jones, M., Cain, B., Mallon, D., and Harris, E. (2018). Rapid isolation and characterization of microsatellites in the critically endangered mountain bongo (*Tragelaphus eurycerus isaaci*). *Journal of Genetics*, 97(2):549–553.

Davey, J., Hohenlohe, P., Etter, P., Boone, J., Catchen, J., and Blaxter, M. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12(7):499–510.

De Barba, M., Miquel, C., Lobréaux, S., Quenette, P., Swenson, J., and Taberlet, P. (2016). High-throughput microsatellite genotyping in ecology: improved accuracy, efficiency, standardization and success with low-quality and degraded DNA. *Molecular Ecology Resources*, 17(3):492–507.

Dieringer, D. and Schlötterer, C. (2003). Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Resources*, 13(10):2242–2251.

Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high-throughput. *Nucleic Acids Research*, 19(32):1792–7.

Ekblom, R. and Wolf, J. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, 7(9):1026–1042.

Fox, G., Darolti, I., Hibbitt, J., Preziosi, R., Fitzpatrick, J., and Rowntree, J. (2018). Genetic assessment of *ex situ* populations to aid species conservation and maintain heterozygosity in non-model species. *Journal of Zoo and Aquarium Research*, 6(2):50–56.

Goldstein, D., Linares, A., Cavalli-Sforz a, L., and Feldman, M. (1995). An evaluation of genetic distances for use with microsatellite loci. *Genetics*, 139(1):463–471.

Goldstein, D. and Pollock, D. (1997). Launching microsatellites: a review of mutation processes and methods of phylogenetic inference. *Journal of Heredity*, 88(5):335–342.

Goldstein, D. and Schlötterer, C. (1999). *Microsatellites: evolution and applications.* Oxford University Press, Oxford, United Kingdom.

Griffiths, S., Fox, G., Briggs, P., Donaldson, I., Hood, S., Richardson, P., Leaver, G., Truelove, N., and Preziosi, R. (2016). A Galaxy-based bioinformatics pipeline for optimised, streamlined microsatellite development from Illumina next-generation sequencing data. *Conservation Genetics Resources*, 8(4):481–486.

Grimaldi, M. and Crouau-Roy, B. (1997). Microsatellite allelic homoplasy due to variable flanking sequences. *Journal of Molecular Evolution*, 44(3):336–40.

Hale, M., Burg, T., and Steeves, T. (2012). Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. *PLOS ONE*, 7(9):e45170.

Hosseinzadeh-Colagar, A., Haghighatnia, M., Amiri, Z., Mohadjerani, M., and Tafrihi, M. (2016). Microsatellite (SSR) amplification by PCR usually led to polymorphic bands: Evidence which shows replication slippage occurs in extend or nascent DNA strands. *Molecular Biology Research Communications*, 5(3):167–174.

Koboldt, D., Steinberg, K., Larson, D., Wilson, R., and Mardis, E. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1):27–38.

Koressaar, T. and Remm, M. (2007). Enhancements and modifications of primer design program primer3. *Bioinformatics*, 15(23):1289–91.

Nichols, J., Conroy, G., Kasinadhuni, N., Lomont, R., and Ogbourne, S. (2018). *In silico* detection of polymorphic microsatellites in the endangered isis tamarind, *Alectryon ramiflorus* (sapindaceae). *Applications in Plant Sciences.*, 6(11):e01196.

Oosterhout, C., Weetman, D., and Hutchinson, W. (2005). Estimation and adjustment of microsatellite null alleles in nonequilibrium populations. *Molecular Ecology Notes*, 6(1).

Price, M. and Hadfield, M. (2014). Population genetics and the effects of a severe bottleneck in an *ex situ* population of critically endangered Hawaiian tree snails. *PLOS ONE*, 9(12):e114377.

Puckett, E. (2016). Variability in total project and per sample genotyping costs under varying study designs including with microsatellites or SNPs to answer conservation genetic questions. *Conservation Genetics Resources*, 9(2):289–304.

Ribout, C., Villers, A., Ruault, S., Bretagnolle, V., Picard, D., Monceau, K., and Gauffre, B. (2019). Fine-scale genetic structure in a high dispersal capacity raptor, the Montagu's harrier (*Circus pygargus*), revealed by a set of novel microsatellite loci. *Genetica*, 147(1):69–78.

Rico, C., Cuesta, J., Drake, P., Macpherson, E., Bernatchez, L., and Marie, A. (2017). Null alleles are ubiquitous at microsatellite loci in the wedge clam (*Donax trunculus*). *PeerJ*, 5(e3188).

Rose, O. and Falush, D. (1998). A threshold size for microsatellite expansion. *Molecular Biology and Evolution*, 15(5):613–615.

Sefc, K., Payne, R., and Sorenson, M. (2003). Microsatellite amplification from museum feather samples: Effects of fragment size and template concentration on genotyping errors. *The Auk*, 120(4):982–989.

Shikano, T., Ramadevi, J., Shimada, Y., and Merilä, J. (2010). Utility of sequenced genomes for microsatellite marker development in non-model organisms: a case study of functionally important genes in nine-spined sticklebacks (*Pungitius pungitius*). *BMC Genomics*, 11(334).

Shin, G., Grimes, S., Lee, H., Lau, B., Xia, L., and Ji, H. (2017). CRISPR-cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis. *Nature Communications*, 8(14291).

Silva, F., Torrezan, G., Brianese, R., Stabellini, R., and Carraro, D. (2017). Pitfalls in genetic testing: a case of a SNP in primer-annealing region leading to allele dropout in BRCA1. *Molecular Genetics and Genomic Medicine*, 5(4):443–447.

Stágel, A., Gyurján, I., Sasvári, Z., Lanteri, S., Ganal, M., and Nagy, I. (2009). Patterns of molecular evolution of microsatellite loci in pepper (*Capsicum spp.*) revealed by allele sequencing. *Plant Systematics and Evolution*, 281(1-4):251–354.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B., Remm, M., and Rozen, S. (2012). Primer3–new capabilities and interfaces. *Nucleic Acids Research*, 40(15):e115.

Vieira, M., Santini, L., Diniz, A., and Munhoz, C. (2016). Microsatellite markers: what they mean and why they are so useful. *Genetics and Molecular Biology*, 39(3):312–328.

Wang, C., Schroeder, K., and Rosenberg, N. (2012). A maximum-likelihood method to correct for allelic dropout in microsatellite data with no replicate genotypes. *Genetics*, 192(2):651–669.

Witzenberger, K. and Hochkirch, A. (2011). *Ex situ* conservation genetics: a review of molecular studies on the genetic consequences of captive breeding programmes for endangered animal species. *Biodiversity and Conservation*, 20(9):1843–1861.

Zhan, L., Paterson, I., Fraser, B., Watson, B., Bradbury, I., Ravindran, P., Reznick, D., Beiko, R., and Bentzen, P. (2016). MEGASAT: automated inference if microsatellite genotypes from sequence data. *Molecular Ecology Resources.*, 17(2):247–256.

3.11 Tables and Figures

Table 3.1

A summary of the design methods used in each species, including the data set number (ID), species, treatment (Tx), number of individuals sequenced (N), number of PCR primers tested (Pp), number and percentage of PCR primers tested successfully amplifying in 75% of samples tested (Amp.), number and percentage of amplifiable PCR primers producing informative data after capillary electrophoresis (easily interpretable and polymorphic) (Inf.), proportion of amplifiable markers which were informative (Inf/Amp), proportion of primer pairs tested which were informative (Inf/Pp), genome size estimate (C-val.), raw sequence reads per sample (mean and SD given where MiMi applied), estimated sequence coverage (Cov.), literature reference and/or accession numbers of NGS data (REF / SRA) where applicable. All genome sizes were retrieved from the Animal Genome Size Database (www.genomesize.com) with the closest related species used. Panels of markers were developed in *P. miliaris* and *C. caeruleus* using both the pal_finder traditional method (Castoe et al., 2015; Griffiths et al., 2016) and MiMi methods. The application of the MiMi quality control process produces higher rates of both amplification and production of informative markers in both these instances.

Table 3.1

ID	Species	Tx	N	Pp	Amp	Inf	Inf/Amp	Inf/Pp	C val	Reads	Cov	REF/SRA
1	*C. caeruleus*	MiMi	8	10	10	8	80%	80%	1.47	8* 2,901,027 (STDEV +/- 878,838)	1.20X	SRX5066864 - 69
2	*P. miliaris*	MiMi	8	20	19	18	95%	90%	1.30	8* 1,482,736 (STDEV +/- 280,686)	0.57X	SRX5162614 - 21
3	*T. eurycerus isaaci*	Trad.	1	30	21	18	86%	60%	3.94	8,980,510	1.10X	(Combe et al., 2018) & SRX5116712
4	*N. pygamaeus*	Trad.	1	30	26	17	65%	57%	3.58	5,309,686	0.74X	SRX5112421
5	*P. miliaris*	Trad.	1	24	13	9	69%	38%	1.30	1,359,615	0.52X	SRX5162614
6	*C. caeruleus*	Trad.	1	10	4	1	25%	10%	1.47	3,913,299	1.60X	SRX5066867

Table 3.2

Species	pal_finder loci	pal_filter loci	MiMi loci
Cyanistes caeruleus	158,147	4,513 (2.9%)	302 (0.19%)
Psammechinus miliaris	469,047	5,657 (1.2%)	250 (0.05%)

Table 3.3

ID	Species	Total	Low Quality Alignments	Primer mutations	Non-variable	High Quality	Good Quality
1	*Cyanistes caeruleus*	302	14 (4.6%)	7 (2.3%)	205 (67.9%)	13 (4.3%)	63 (20.9%)
2	*Psammechinus miliaris*	250	102 (40.8%)	9 (3.6%)	101 (40.4%)	12 (4.8%)	26 (10.4%)

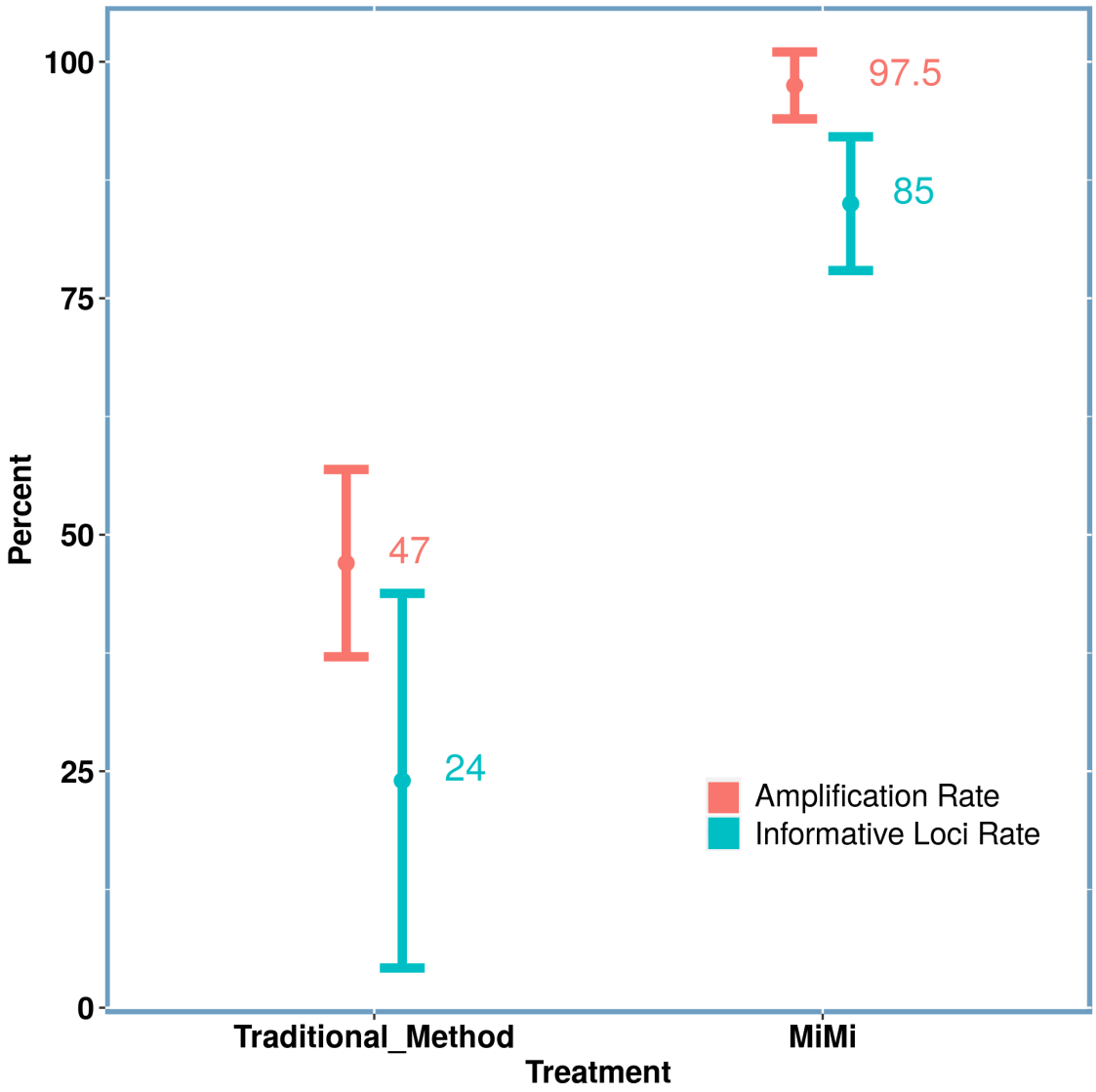

100
75
50
25
0
Percent
97.5
85
47
24
Amplification Rate
Informative Loci Rate
Traditional_Method
MiMi
Treatment

Figure 3.3

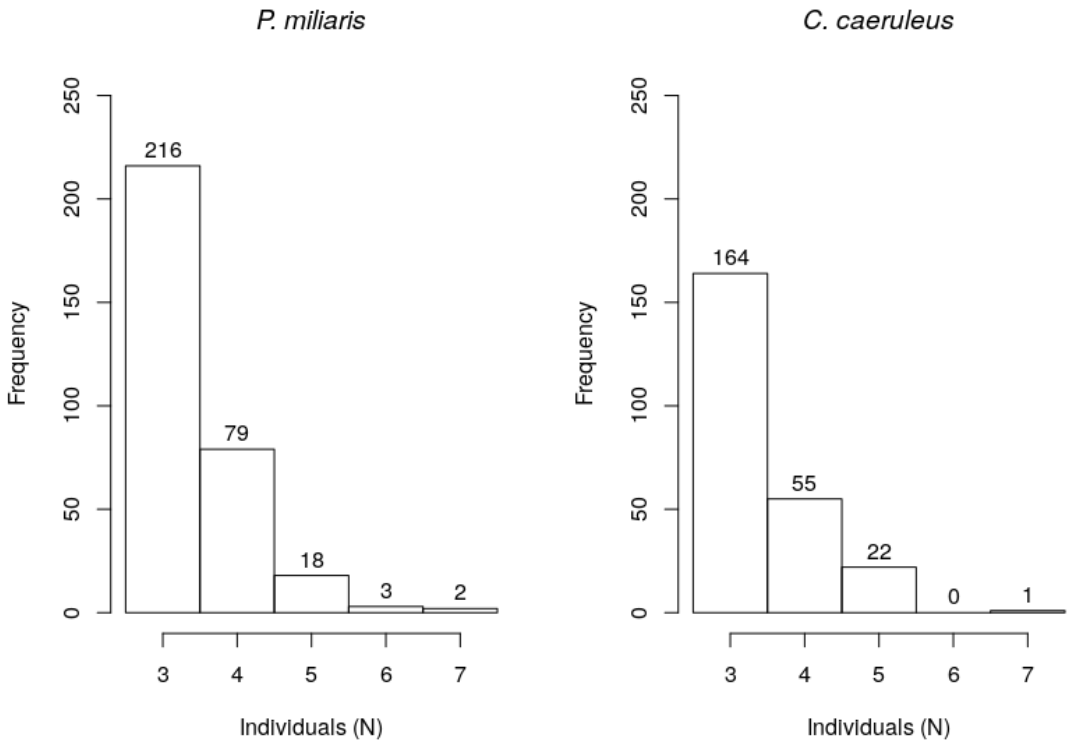

P. miliaris
Frequency
0
50
100
150
200
250
216
79
18
3
2
3
4
5
6
7
Individuals (N)
C. caeruleus
Frequency
0
50
100
150
200
250
164
55
22
0
1
3
4
5
6
7
Individuals (N)

Figure 3.4

Chapter 4

A Comparative Study into the Power and Application of Microsatellites and High-Throughput SNPs for Population Genetics.

4.1 Effective genetic markers for population structure analysis of the larval dispersing decapod, *Homarus gammarus* (the European lobster).

Graeme Fox [1], Richard F. Preziosi [1] and Jennifer K. Rowntree [1]

[1]Ecology and Environment Research Centre, Department of Natural Sciences, Manchester Metropolitan University, Chester Street, Manchester, United Kingdom, M1 5GD

Keywords: single-nucleotide polymorphisms, microsatellite, comparison, population genetic structure, lobster, *Homarus gammarus*.

Author contributions: GF, RFP and JKR conceived and designed the project; GF sourced and coordinated the sample collection; GF performed the lab work; GF performed the data analysis; GF wrote the chapter.

4.2 Abstract

Globally, fishing resources are under unprecedented pressure to meet the food demands of the ever-increasing human population. *Homarus gammarus* (the European lobster) fisheries in Europe have been a long-lived, highly profitable source of sustenance throughout their range, however in the latter part of the 20th century dramatic stock collapses occurred in some regions. Stock conservation efforts in the UK and Ireland are focused on preventative measures and also supportive stocking from several hatcheries. The aim of this study was two fold. Firstly, we explored the extent to which the local hydrodynamics around the UK and Ireland allow genetic mixing between the different regions, enabled by the relatively long larval life stage of *H. gammarus*. Secondly, we performed a comparative study of the ease of application, cost-to-benefit ratio, and relative power of two types of genetic marker typically used for population genetic analysis; microsatellites and single nucleotide polymorphisms (SNPs). In ecological genetics and conservation, there is a sustained and significant move away from the use of microsatellites, to the use of SNP markers in population genetics. We performed parallel analyses using both types of marker to compare their resolution and test their suitability for a study of this nature. We found an absence of detectable population structuring using either marker type indicating high levels of population mixing and therefore low population isolation by genetic distance. Comparisons in the practicability and cost-to-benefit of each marker reveal that SNP analysis is technically more expensive, but much quicker approach for population genetic analysis. In our case to generate comparable results, SNP analysis was performed in just a few weeks, whilst microsatellite analysis took many months, as marker and multiplex optimisation were required. Our results are informative for conservationists interested in the marker choice for their study, and is the first comparative study of this type in a larval-dispersing organism.

4.3 Introduction

4.3.1 Molecular Markers for Population Genetics

Microsatellites have been used for decades as molecular markers in studies of population genetics (Vieira et al., 2016), but increasingly are being superseded by single nucleotide polymorphism (SNP) markers (Helyar et al., 2011). The development of

high-throughput sequencing technologies has surmounted many of the difficulties and expenses associated with analysis of a large number of SNPs, and well documented methodologies have brought them within reach of many ecology laboratories (Catchen et al., 2013; Fischer et al., 2017). Microsatellites require *a priori* knowledge of the genome of the study species, as PCR primers must be designed to amplify the marker regions. However, SNP markers can be analysed using a completely *de novo* approach (Benestan et al., 2015), free from the design of PCR primers, with clear benefits to species with little previous study. The trade off in relative power of each individual locus (the multi-allelic microsatellite locus vs. the bi-allelic SNP) is countered by the number of individual marker loci which can be practically, concurrently analysed (typically 10-20 microsatellites or thousands of SNPs), with increasing marker number correlated with greater power to detect genetic structure (Coates et al., 2009; Lemopoulos et al., 2019; Souza et al., 2019; Gkagkavouzis et al., 2019; Gärke et al., 2012). Analysis of the variation at a relatively small number of microsatellite loci is generally assumed to constitute a representative sample of wider genomic variation (Selkoe and Toonen, 2006), however in the few studies that have performed direct comparisons between microsatellites and SNPs, which are more widespread through the genome, this has not always been the case (Roques et al., 2019; Coates et al., 2009).

The relative power of a panel of microsatellites and a panel of SNP markers to determine genetic population structure has been investigated in several species. A study into the population genetics of the bowhead whale (*Balaena mysticetus*) discovered that 42 SNP markers gave a similar level of power to calculate F_{ST}, perform population assignment and to estimate effective population size (N_e) as a panel of 22 multi-allelic microsatellite markers, in a population with low levels of genetic differentiation (Morin et al., 2012). A multiplex of 18 microsatellites was found to be significantly less powerful, than 79 SNPs for parentage assignment in the European sea sturgeon (*Acipenser sturio*), with maximum SNP resolution achieved with just 28 SNP loci (Roques et al., 2019). Both these studies utilised relatively few SNP markers, and performed genotyping using microarray type technologies. Using high-throughput methods of genotyping 481 restriction site associated (RAD) derived SNPs, and drawing comparisons to 32 microsatellites, a study into the mangrove killifish (*Kryptolebias marmoratus*) found that whilst both datasets gave similar results, the SNP data suffered from a high-rate of signal noise, limiting resolution, highlighting the associated requirements for caution when inferring fine-scale population structure and the

importance of quality checking routines (Mesak et al., 2014). The utility of both SNPs and microsatellites is clear, with different marker types offering different benefits. Further direct comparisons using current sequencing and genotyping methods, in species with a range of life histories will further illuminate the apparent progression from microsatellites to SNPs, and be informative to ecologists planning studies in population genetics (Vignal et al., 2002; Luikart et al., 2003).

Variation in electrophoretic rates during capillary electrophoresis, and human or algorithmic error introduced during peak-calling analysis can all cause significant error rates in microsatellite analysis, and make cross compatibility between labs extremely difficult (Pasqualotto et al., 2007; Alberto, 2009). Analysis of SNPs is concerned with categorical, binary variation at each SNP site and alleles are much less ambiguous (Helyar et al., 2011). Massively parallel analysis of samples for SNP analysis can occur on a single NGS run, meaning that technical variation in data generation is less of a concern, and laboratory cross compatibility possible. A critical factor in SNP analysis is the depth of sequencing coverage achieved at each SNP site. Depth of sequencing coverage is determined by several factors which may be unknown prior to sequencing: data output of the sequencer; sample size; genome size; frequency of restriction site occurrences in the genome; and frequency of SNPs close to restriction site (Catchen et al., 2013). These factors should be carefully considered during experimental design.

Ultimately, marker choice is likely to be study dependent with one marker type very unlikely to consistently provide the optimum approach. Significant factors in their comparable effectiveness will likely be the degree of population differentiation in the species of interest, sample number, potential of concurrent analysis of samples, and budget availability. Comparative studies between different marker types for population genetics contribute a frame of reference which may inform future studies in taxonomically similar species, in species with a similar mode of distribution or life cycle, or species with a similar putative population structure.

4.3.2 Conservation of *Homarus gammarus* Fisheries in the UK and Ireland

The European lobster (*Homarus gammarus*) is a widely distributed, decapod crustacean found throughout areas with rocky substrate in north west Europe and north west Africa, extending into the Mediterranean Sea (Holthuis, 1991; Cobb and Castro, 2006). In 2017 (latest published figures), the fishing industry in the United Kingdom

(UK) supported around 12,000 fishers and had the second largest catch of the European Union (EU) countries, landing roughly 724,000 tonnes per annum. Shellfish catch in the UK has seen a proportional decline since 2007, but remains the largest share of total landings at 38% (the remainder of the catch being either pelagic or demersal landings) and is increasing year on year in live weight value (ca. £2500 tonne^{-1} in 2017). Similarly, *H. gammarus* is in the top 20 species of greatest value to the total catch of Ireland in 2017, and is the single most valuable species caught by the UK fleet at £13 Kg^{-1} (Elliott and Holden, 2017). The *H. gammarus* catch landed in the UK was reasonably consistent in the period 2013-2017 (ca. 324,000 tonnes per annum, STDEV: +/- 18,000 tonnes) but has increased in value by approximately 50% in the same time period. In 2017, the lobster was the second most valuable species per tonne in Ireland (Elliott and Holden, 2017; Sea-Fisheries Protection Authority, 2017). Lobster is a particularly valuable shellfish catch due to its status as a high-end seafood product and associated high market value (Ingebrigt et al., 2005). Lobster fisheries are often extremely important to local economies supporting fishers, restaurateurs and enhancing appeal to tourists (Brookfield et al., 2005; Browne et al., 2001). In the context of devastating stock collapses in Europe over the last half-century (Kleiven et al., 2018), considerable effort has been made in the UK and Ireland to preserve remaining stocks through the implementation of legislation to limit exploitation (European Commission, 2014) and through several supportive breeding and restocking programs.

There are many reasons why genetic surveys can be beneficial for conservation. One of the major threats to global biodiversity is the fragmentation of habitat, and associated increasing numbers of spatially isolated populations (Saunders et al., 1990). Prolonged isolation of fragmented populations can result in an increase in genetic drift, increased inbreeding, and greater likelihood of the loss of local adaptations. These populations are at increased risk as smaller, genetically distinct populations tend to be less persistent due to losses of heterozygosity and population viability (Young et al., 1996). Genetic markers can be used to detect and measure the rate of gene flow between different geographic sites, and therefore imply that migration, or breeding is occurring between and amongst study sites. By providing this information, genetic data can be used to identify genetic stocks, or management units, which can be used to ensure that managed exploitation of resources allows the restoration of the stocks at rates above that of the maximum sustainable yield (Casey et al., 2016). Data can also be used in order to determine adaptive divergence of populations. Populations should be adapted to their localised environments,

however for divergence to occur, there needs to be sufficient isolation or the selective pressure of a particular subpopulation needs to be sufficiently strong to exceed the effects of migration. Adaptive divergence can often be detected by a subset of genetic markers, often SNP markers, which rather than being selectively-neutral, are associated with ecologically relevant traits (Wagner et al., 2017), and can exhibit rapid adaptation (Willoughby et al., 2018). Loss of localised adaptation through over-exploitation, or other ecological pressures, can leave natural populations with a higher probability of population collapse or extinction.

Information relating to connectivity of populations and levels of self-recruitment benefit the conservation of complex natural resources, such as fishing stocks (Hastings and Botsford, 2006), including the identification and management of marine protected areas (MPAs), (Watson et al., 2016). Measuring genetic variation within and among populations is extremely informative to the conservation of a natural resource, particularly when an activity such as harvesting, or population augmentation is taking place (Lemopoulos et al., 2019). The accurate identification of evolutionarily significant units (ESUs) is critical where calculations relating to harvest and recovery are made by conservation authorities. Where a subpopulation is distinct from others through genetic isolation or other forms of adaptation, each separate population may require a different management strategy (Frankham et al., 2004).

The rate of maximum sustainable yield (MSY) of a population is highly informative to conservation strategy. The calculation of MSY allows conservationists to determine the rate of harvest which maximises the population recovery rate, controlling the risk of over-exploitation. Effective population size (N_e) affects genetic variation and its distribution on geographic and temporal scales. Conversely, the use of genetic data to quantify the extent of genetic variation can therefore be used to estimate N_e (Wang, 2005). Effective population size can be estimated from genetic data by a number of methods. For example measuring heterozygote excess at multiallelic loci (such as a microsatellite), and through the analysis of the general relationship between N_e and heterozygosity (Pudovkin et al., 1996; Luikart and Cornuet, 1999; Wang, 2009). Furthermore, N_e can also be estimated from measures of linkage disequilibrium recorded in genetic data at multiple polymorphic loci (Hill and Robertson, 1968; Hill, 2009). Whilst the effective population size and MSY are clearly linked, the MSY does not increase linearly with an increasing population size. Treating a natural biological resource as a single large population and harvesting at or around the MSY, where it is actually several smaller distinct populations, would result in over-exploitation and

severely limit the recovery of the population. An accurate determination of the population structure is therefore critical if appropriate management strategies are to be used effectively (Allendorf et al., 2012).

In benthic species with a relatively long-lived, planktonic larval life stage, such as *H. gammarus*, dispersal and recruitment often occur at a broad scale in the meta-population, with wide-spread mixing of sub-populations dictated largely by coastal currents (North et al., 2008). After the larval life stage, which can last from three to ten weeks depending on conditions such as water temperature, the animal adopts a benthic life style which is retained throughout adulthood, with only very limited migration (Rötzer and Haug, 2015; Werner, 2007). Isolated populations that do not benefit from this widespread mixing and recruitment, are at additional risk of over-exploitation and population collapse if they are incorporated into a broader conservation strategy assuming a meta-population with high genetic mixing (Watson et al., 2016). Management of a species with an unclear dispersal range, such as those with a free-living planktonic dispersal mechanism, can be extremely difficult due to limited knowledge of the extent of links between external influencing factors and dispersal range. Population structure analysis techniques using genetic markers are therefore an invaluable resource to provide information relating to population differentiation and dispersal (Watson et al., 2016).

Molecular analyses of several areas of the *H. gammarus* range have been performed previously. There has been an almost complete absence of population structuring detected in the region around the UK and Ireland (Watson et al., 2016) using microsatellite analysis, but some moderate structuring apparent at a broader scale, when genotypes from the British archipelago were compared to other European regions, notably the Skagerrak and Kattegat off the Swedish and Danish coasts (Ellis et al., 2017). A recent assessment of the broader European population using SNP markers described a genetic cline across the European region, supporting evidence from other lobster species that population structure may be present, albeit potentially on a large scale or as patchy structure not detectable by studies upon a limited range (Jenkins et al., 2019; Truelove et al., 2013). Given the long-lived planktonic dispersal of the species, it is possible that genetic mixing is at such a high level that for conservation purposes, much of the western European population should be treated as a single panmictic conservation unit. However, it may also be possible that the absence of any detectable population structuring may represent the technological limitations of the analysis methods which have so far been implemented.

Related studies have been performed on other marine species, with a similar mode of dispersal, with mixed results. For example, an analysis of the genetic population structure of the great scallop (*Pecten maximus*), which has a larval period very similar to that of *H. gammarus*, was unable to detect any population structuring along the coast of Northern Ireland using a panel of 13 microsatellites, however the increased power and resolution afforded by a screen of 10,539 SNPs allowed the detection of several significant pairwise F_{ST} values amongst the sites, and lead to the identification of two genetic clusters (Vendrami et al., 2017). A microsatellite based analysis of the genetic population structure of the brown crab (*Cancer pagurus*) found an absence of genetic structure within the Irish Sea or at a regional level, and a similar result was found using an allozyme-based analysis in the Norway lobster (*Nephrops norvegicus*), sampled in the Eastern Atlantic and Mediterranean sea (McKeown et al., 2018; Stamatis et al., 2006).

Dispersal in *H. gammarus* is driven by the relatively extended period of planktonic, larval development. The development consists of several distinct larval life stages, three of which are characterised by an omnivorous, pelagic period during which they float freely in the surface layers of the ocean, for a period of approximately three to ten weeks, before settling into a benthic life style, with very limited further migration as an adult (Rötzer and Haug, 2015). This single, time-limited opportunity for dispersal is dictated largely by currents, with the larvae able to direct themselves modestly in the water column, mostly in on the vertical axis (Schmalenback and Buchholz, 2010). Despite this, some larval dispersers (including lobsters) do exhibit some population structuring, apparently violating the assumption that such a capacity for very broad dispersal would lead to panmixia (Babbucci et al., 2010).

Larvae are released as eggs hatch during the summer months (Pandian, 1970), when the currents around the UK and Ireland are dictated largely by the North Atlantic current. The current arrives at the South West of the region, passes along the Western coast of Ireland and Northern Ireland, and runs north through the Irish Sea, before turning East into the North Sea. Limited mixing from the North Sea, back into the Irish Sea occurs via the English Channel due to a current in the channel which mostly runs West to East but also the presence of a series of gyres along the coast. Finally, a near-surface gyre in the Irish Sea is present in the summer which may facilitate larval retention in the region (Hill et al., 1997; Taylor, 1995; Zheng et al., 2002; Ménesguen and Gohin, 2006). Based upon these broad observations of seasonal currents and gyres around the UK and Ireland, our hypothesis regarding any potential genetic structure in *H. gammarus* is that we are

most likely to see distinct genetic groups between sites on the West of Ireland and those on the coast of Cornwall and the Irish Sea. The most mixed sample sites are likely to be those on the east coast of Scotland and Northern England, as those likely receive larvae from majority of the other sample sites. Gyre mixing may cause some retention of larvae in the Irish Sea, isolating those sites from moving north and ultimately mixing with sites in the east.

Our joint aims were to describe any detectable genetic structure in the samples, and to compare the results from two types of molecular marker typically used in population genetics: microsatellites and SNPs. There have been limited tests or direct comparisons between these two approaches generally (Lemopoulos et al., 2019; Jeffries et al., 2016; Morin et al., 2012; Roques et al., 2019; Mesak et al., 2014), and none to our knowledge performing such a comparison in a species with a long-lived planktonic larval stage. Towards these goals, we have performed parallel population genetic analysis of *H. gammarus* samples from around the coasts of the UK and Ireland using both types of genetic marker and performed statistical analysis of the genotypes to assess the presence of any genetic structure. Furthermore, we present practical comparisons of the two methods, taking work-load and cost-benefit into account and present our recommendations for future population genetics work on this, and similar species.

4.4 Materials and Methods

Samples were collected between 2014 - 2017 from 15 sites around the United Kingdom and Ireland in accordance with local fishing regulations and restrictions. Samples were collected by commercial fishers coordinated by either a local fishing conservation authority, conservation trust, academic or hatchery (Table 4.1 & Figure 4.1).

Wild *H. gammarus* were caught using lobster pots and a small section of pleopod removed using clean scissors and forceps. The tissue sample was placed immediately into 100% ethanol for preservation. Upon receipt of samples at Manchester Metropolitan University, the tubes were stored at -80°C until further processing. Total DNA was extracted from approximately 25mg pleopod tissue sample using the Wizard® SV 96 DNA purification system (Promega, Wisconsin, USA) according to the manufacturers protocol except that the lysis incubation was carried out overnight to ensure complete tissue lysis. Purified DNA was checked for quality and quantity on a NanoDrop 3000 spectrophotometer (Thermo Fisher Scientific, Massachusetts, USA) and a 1% agarose

electrophoresis gel. Samples were normalised to 20 ng/µl where the elute was originally at a higher concentration.

4.4.1 Microsatellite Development and Analysis

Fourteen published microsatellite markers (André and Knutsen, 2010; Ellis et al., 2015) and six novel tetra-nucleotide markers mined from sequence data provided by Dr. Charlie Ellis (National Lobster Hatchery, UK) were tested in the laboratory. Novel markers were developed using a Galaxy (Afgan et al., 2018) bioinformatics methodology (Griffiths et al., 2016), modified to handle single-end, Sanger sequencing reads (Table 4.2). Markers were amplified using six multiplexes, using a universal tail PCR approach (Blacket et al., 2012), (Table 4.3). Three fluorophores were used (6-FAM, TAMRA and PET) to fluorescently label PCR products. Reactions were performed using the Type-it Microsatellite PCR kit (Qiagen, Hilden, Germany). Total reaction volume was 5µl consisting of 2.5µl Type-it 2x mastermix, 1.5µl molecular biology grade H_2O, 0.5µl primer mix (2µM stock concentration) and 0.5µl template DNA (normalised to 20 ng/µl). Amplifications were performed using a Techne Prime thermal cycler (Techne, Minnesota, USA). Cycling conditions were as follows: 95°C - 5 minutes; 35x {95°C - 30 seconds; 58-60°C - 90 seconds; 72°C - 30 seconds}; 72°C - 30 minutes; 4°C - Hold. Annealing temperature was dependant on multiplex (Table 4.3). Microsatellite PCR products were run on an ABI 3730 DNA Analyzer (Thermo Fisher Scientific, Massachusetts, USA) at the University of Manchester DNA Sequencing Facility. Raw fragment sizes were determined using the R (Team, 2017) package,"Fragman" (Covarrubias-Pazaran et al., 2016) and alleles binned using the R package "MsatAllele" (Alberto, 2009). Samples which failed to amplify, produced peaks which were not clearly interpretable or otherwise were missing >50% data were removed (58 samples). Similarly, markers which failed to amplify in >33% of samples were removed (2 markers). A random subset of 12% of the samples were repeat genotyped to estimate the error rate in our allele scoring procedures.

4.4.2 RAD-Seq Library Preparation

Ninety five samples, which had previously been analysed using microsatellites, were chosen for SNP analysis, encompassing four regions of the UK and Ireland (Table 4.1). Fresh DNA extractions were performed using same method as previously, extracts normalised to 3.1ng /µl and fragmented with a Covaris M220 sonicator with Covaris

microTUBE AFA Fiber Snap-Cap tubes (Covaris, Massachusetts, USA), with a program optimised for production of an average fragment length of 800bp. A subset of sheared samples were checked for quality using a high sensitivity DNA bioanalyzer chip (Agilent, Santa Clara, CA, USA) before being blunt-ended and A-tailed using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, Hitchin, UK). Ligation reagents from the same kit were used to ligate adapters containing the Illumina Nextera transposase read 2 adapter sequence onto the sonicated DNA, and excess adapter removed using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA). Samples were incubated with the restriction enzyme SbfI (cutsite: TGCA–GG; New England Biolabs, Hitchin, UK) at 37°C overnight to ensure complete digestion, prior to purification with Agencourt AMPure XP beads. Phased SbfI specific adapters containing the Illumina Nextera transposase read 1 adapter sequence were ligated onto sonicated DNA using T4 ligase (New England Biolabs, Hitchin, UK), and excess adapter was removed using Agencourt AMPure XP beads (Wu et al., 2015). Phased adapters were biotinylated at the 5′ on the top strand to allow library enrichment using Dynabeads MyOne Streptavidin C1 beads (Invitrogen, Carlsbad, CA, USA). Samples were then amplified using the Nextera XT Index v2 (Illumina, San Diego, CA, USA) primers and the NEBNext Ultra II Q5 Master Mix PCR reagents (New England Biolabs, Hitchin, UK) to uniquely index each sample, and add sequences required for Illumina sequencing. Library quality of a subset of samples was confirmed using a high sensitivity DNA bioanalyzer chip and all samples quantified using the Qubit dsDNA BR Assay Kit and a Qubit 2.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA), prior to normalisation and preparation to load onto the flowcell.

4.4.3 RAD-Seq Sequence Analysis

Raw sequencing reads were demultiplexed into individual samples, trimmed at the 5′ up to the SbfI cut site, processed with Trimmomatic (Bolger et al., 2014), (LEADING:3, TRAILING:3, SLIDINGWINDOW:4:10, MINLEN:100) and trimmed at the 3′ to a fixed length of 100bp. The STACKS software (Catchen et al., 2013) was used to derive a panel of SNP markers, following a published protocol for optimisation of the parameter space (Paris et al., 2017). Optimised parameters were used to generate a list of SNPs and generate files suitable for downstream analysis. Parameters were: m=2, M=1 and n=2, where m is the minimum number of reads which must be met before a stack (allele) is generated, M is the maximum amount of nucleotide mismatches between stacks for merging into a single

locus, and n is the maximum amount of nucleotide mismatches between stacks allowed during construction of the reference catalog.

4.4.4 Statistical Analysis

All statistical analyses were performed in the R environment (Team, 2017), using the packages described below, unless otherwise stated. In both data sets, observed heterozygosity (H_{OBS}) and expected heterozygosity (H_{EXP}) were calculated using "adegenet" (Jombart, 2008; Jombart and Ahmed, 2011). Likelihood of each pair of loci exhibiting linkage disequilibrium was calculated using "genepop" (Rousset, 2008) and deviation from Hardy-Weinberg equilibrium (HWE) calculated with "pegas" (Paradis, 2010), both with false discovery rate under multiple testing corrected using the "B-Y" method (Benjamini and Yekutieli, 2001). The frequency of null alleles at each locus was estimated using the standalone software "FreeNA" (Chapuis and Estoup, 2007; Chapuis et al., 2008) and the implementation of the EM algorithm (Dempster et al., 1977). F-statistics and rarefied allelic richness (AR) were calculated using "hierfstat" (Goudet, 2005) and rarefied private allelic richness (PAR) calculated with ADZE (Szpiech et al., 2008), with both AR and PAR rarefied to the lowest sample size (rarefied against the eight Boscastle samples). Jost's D estimator of population differentiation was calculated by the "DEMEtics" package (Gerlach et al., 2010). F_{ST} was calculated to attempt to understand the degree of variation attributable to putative population structure, whilst D is a more direct measure of the genetic variation present in the data (Jost, 2009).

In the microsatellite data only, the putative number of population clusters (K) required for Structure (v.2.3.4) analysis (Pritchard et al., 2000) was determined using the "Cluster Identification Using Successive K-Means" method using the "adegenet" package; a method independent of geographic information that uses calculations of the Bayesian Information Criterion (BIC) for each estimate of K until the optimum model is found. Structure analysis (Pritchard et al., 2000) of the microsatellite data was performed using an initial burn in length of 5,000 cycles, followed by 500,000 markov chain monte carlo cycles, using the admixture model, and using both correlated and independent allele frequencies. Analysis was performed at every value for K from 1 to 20.

4.5 Results

4.5.1 Genetic Marker Development

Six new microsatellite markers were developed and optimised from Sanger sequencing reads (Table 4.2).

After the removal of data pertaining to negative controls and four failed samples, the next-generation sequencing of 91 RAD-Seq samples produced 38,132,997 paired-end, 2*150bp raw sequencing reads (mean: 419,043, 95% CI: +/- 103,881.8). These were filtered by QC processes to 8,089,314 reads (mean: 90,812.5, 95% CI: +/-19,986.9). The construction of stacks and detection of SNPs in the population created 597 RAD loci, consisting of 898 SNPs, for use in further analysis.

4.5.2 Microsatellite Analysis

Using 20 microsatellite markers (14 published markers and six novel markers) we were able to genotype 389 wild *H. gammarus* samples from 15 sites (Figure 4.1), (mean sample number: 25.5 per site, SE: 3.75). After removal of samples and markers which failed to amplify above the previously specified thresholds we retained 325 (85%) samples and 18 (90%: HGD117 and HGD129 removed) markers, successfully obtaining 90% genotypic data in this filtered data set. The estimated error rate from this subset was 1.74%, based on inconsistent genotype calls in the random 12% of the data which was genotyped multiple times. Every sampling site was found to contain absolute private alleles with the exception of County Clare and North Shields, however the number of private alleles was found to be strongly correlated with the number of samples collected at each site (Welch two sample paired t-test: t=7.62, df=13, pval<0.0001). To handle variation in per-site sample number, allelic richness and private allelic richness values were rarefied to the lowest sample number (North Shields, N=10), (Table 4.5). Average rarefied allelic richness was 2.212 (SE: 0.008) and average rarefied number of private alleles was 0.132 (SE: 0.013). No pair of markers were found to exhibit significant linkage disequilibrium after B-Y correction for multiple tests (Benjamini and Yekutieli, 2001). The number of observed alleles per locus ranged from six (HGC6) to 22 (GF_13). Overall we saw a significant difference between observed (H_{OBS}) and expected (H_{EXP}) heterozygosity (Bartlett's test of homogeneity of variances: t=2.5937, df=17, pval=0.02), but no single site showed significant variation between H_{EXP} and H_{OBS}. There were 15 incidences where a marker deviated significantly from Hardy-

Weinberg equilibrium (HWE) within a sampling site, after B-Y correction (Table 4.6), but the only site with a large number of markers not in HWE was Firth of Forth with eight. On average no marker had a high (>10%) estimated frequency of null alleles (mean = 0.02, SD = 0.02), however there were 21 instances where a single marker showed high evidence of null alleles in a single population (Table 4.7).

Global F_{ST} was used as an estimator of population differentiation and was reported to be 0.002 indicating that very little variation in the genotype data can be explained by putative population structure. Pairwise F_{ST} among populations ranged from 0.005 (little genetic differentiation) between the sites at Amble and Firth of Forth to 0.098 (moderate genetic differentiation) between Boscastle and North Shields (Hartl and Clark, 1997). Jost's measure of population differentiation showed no significant results after p-value correction for any pair of sites (Table 4.8). Determination of K using "adegenet" was unable to give a definitive results, with three, four or five clusters appearing equally likely. Structure analysis at these values for K indicated no describable population structure (Figure 4.2). Given the apparent absence of detectable population structuring by microsatellites among our individual sampling sites, we also classified samples into one of four broad sampling sites (Table 4.1). This effective increase in per-site sample number should increase the statistical power of the data, however global F_{ST} between these broader putative populations remained very low (F_{ST}=0.002), and pairwise F_{ST} was <0.007 in every instance (Table 4.9).

4.5.3 SNP Analysis

Using the 579 SNP markers designed by the STACKS process (Catchen et al., 2013), we were able to genotype 91 samples, using markers which were present in ≥40% of samples. The broader geographical sampling sites (described previously) were used for SNP analysis, as the sequencing did not provide us with sufficient depth of coverage to develop and genotype SNP markers at the finer geographic scale. Significant deviation from HWE was seen in 74 markers, and linkage disequilibrium was detected to be significant in 2 pairs of markers, which were all subsequently removed from further analysis. Observed heterozygosity was significantly greater than expected heterozygosity in all putative populations (Table 4.10). PHRED quality scores in the data were good, with any very low regions (<10) removed previously. Aligning sequence data into stacks covering the same region allows even regions with relatively low individual quality scores to be improved, by the consensus of several reads. After quality control, per-site sample

numbers ranged from 21-23, and raw reads per-site from 1.4 million to 3.0 million (North East: mean=65,708.64, SD +/- 9,149.45; South West: mean=66,395.67, SD +/- 18,869.87; Atlantic Ireland: mean=131,192.3, SD +/- 21,055.43; Irish Sea: mean=96,738.78, SD +/- 25,813.99), (Table 4.10). Average rarefied allelic richness was 31.810 (SD +/- 11.872) and private allele allelic richness was 118.108 (SD +/- 98.908). Both rarefied allelic richness and private allelic richness were much greater in the Atlantic Ireland site, compared to the other three sites, likely linked to the greater amount of sequence data generated for that site (Table 4.10).

To test for population structuring, global F_{ST} was calculated and was found to be extremely low (0.005), indicating an absence of detectable population structuring. Pairwise F_{ST} and D calculations support this, with every pair of populations reporting values of approximately zero (Table 4.9).

4.6 Discussion

Our first aim was to use molecular markers to investigate the population structure of the wild *Homarus gammarus* population(s) around the coast of the UK and Ireland, and to use our findings to inform conservation strategy of this economically important species. Of the two genetic data sets we analysed, one derived from microsatellite genotypes and one derived from SNP genotypes, neither panel of markers were sufficiently statistically powerful to adequately differentiate any genetic clusters within the data. We detected very low F_{ST} and D values between almost all pairs of sites, at both geographic resolutions, indicating an almost complete absence of any detectable genetic structure. During analysis of the microsatellite data set only, pairwise F_{ST} between Boscastle and three other sites (County Clare, North Shields and Waterford) showed some moderate genetic differentiation. This may suggest some level of genetic isolation, but is likely linked to the small sample number sourced from Boscastle, with small sample sizes known to cause over-estimation of F_{ST} (Willing et al., 2012). This moderate genetic differentiation was not recorded in the analysis of broader geographical sites, or with the SNP marker data which is more suited to smaller sample numbers, suggesting this is likely an anomalous result (Chapuis and Estoup, 2007).

4.6.1 *Homarus gammarus* population structure in the UK and Ireland

In both data sets we saw high levels of genetic variation within every sample site, evidenced by high total allelic number and the presence of private alleles in most sites. Private alleles can be indicative of low levels of genetic mixing, allowing a population to retain an allele in this fashion, or may be detected in a genetic screen as a result of undersampling. Given the overall high rates of genetic mixing detected in this dataset, and correlation between number of private alleles, and number of samples-per-site, we would suggest that the private alleles detected in this instance are a result of undersampling.

Both SNP and microsatellite data showed high heterozygosity and overall diversity, in spite of intense harvesting, indicating a large evolutionary potential and low levels of inbreeding at present (Frankham et al., 2004). All markers were highly variable, with previous simulation studies on 12 of the microsatellite markers used here, suggesting high capacity to differentiate genetic structure (Watson et al., 2016). Individual microsatellite markers were very infrequently out of Hardy-Weinberg Equilibrium (HWE) within, or among sites, indicating high levels of genetic mixing, generally associated with a panmictic population (Gillespie, 1998). There was some evidence of null alleles in the microsatellite data set. The presence of null alleles at a locus can cause an increase in the rate of homozygotes, where a single allele fails to amplify, or of missing data where both alleles fail to amplify (in diploid species). A reduction in the rate of heterozygotes, generally through one or more primers failing to bind due to mutations in the primer binding site (Rico et al., 2017) can lead to over emphasis on deviation from HWE and F-statistics. The highest per-marker estimate of the frequency of null alleles across all samples was 0.07, with just 21 individual site and microsatellite marker combinations showing relatively high estimated rates of null alleles. Given the low values of deviation from HWE and F_{ST}, these estimates of relatively infrequent null alleles are unlikely to bear significant impact upon our results, which already show low levels of genetic differentiation.

A common issue with microsatellite genotyping for population genetics is the under-estimation of measures of differentiation, including F_{ST} and G_{ST}. This phenomena is caused by the highly polymorphic nature of microsatellites and their rapid evolution; namely that even in two completely genetically distinct populations, you will very likely find overlapping alleles which arose due to evolutionary constraints upon repeat number, or back mutations. Statistical models built upon the quantification and comparison of

allelic variation within a sub-population to the total population do not require the identification of the alleles themselves. In microsatellite datasets, where two distinct populations may have co-evolved the same alleles (homoplasy) at the locus, this can cause an order-of-magnitude depression of differentiation measures (Hedrick, 1999; Balloux et al., 2000). Whilst our microsatellite dataset may well be suffering from this depression of differentiation metrics, we found strong correlation with differentiation statistics generated by the SNP dataset, and with other work on the species in the region (Watson et al., 2016) and as such do not consider this factor to be a major concern.

The very low values of F_{ST} detected could be indicative of a population which has not yet reached equilibrium after a recent disturbance, and is in the process of population differentiation, ultimately potentially leading to vicariance (Hoorn et al., 2010; Smith et al., 2014; Barreiro et al., 2008). In an evolutionary timeframe, a large portion of the UK and Ireland was recently covered with an extensive ice sheet stretching from SW Ireland to NE Scotland covering Kerry, the English Lake District and Wales. The sheet peaked in its extent at 27 ka BP (kilo annum before present) and retreated north over the next 12 ka, resulting in the scenario we have today, which has existed for approximately the last 15,000 years (Clark et al., 2010). Given the rapid, broad dispersal capability of *H. gammarus*, we can assume a relatively rapid colonisation of the northern coasts of the region following the receding ice sheet as the southern and continental populations spread north into the new territory. A hypothesis could therefore be argued that western and northern populations of Europe are in the process of differentiation, however there has not yet been sufficient generations in this slow growing species for a signal to yet be detectable by these genetic markers (Rötzer and Haug, 2015; Werner, 2007; Jenkins et al., 2019).

Sampling of wild populations took place over a period of at least three years, but potentially up to five (some archived samples were used for which sampling dates were not available), and as such there is scope for allele frequencies to have varied within each sampling site, and in the population as a whole, year on year. Furthermore there is the possibility of observing annual changes in allele frequencies as a result of sampling error. Harvesting is targetted towards mature, male individuals, whilst the females are actively avoided and in *H. gammarus* we have a long-lived, late maturing species of overlapping generations which in the UK and Ireland has not suffered a historic and dramatic population bottleneck from over-fishing. Allele frequencies are therefore not expected to vary annually, however further sampling and analysis to fully investigate annual fluctuations would be illuminating (McPherson et al., 2011; King et al., 1987). Our

findings are in keeping with several, related studies into larval dispersing marine organisms in the region. Studies investigating *P. maximus* and *C. pagurus* in the Irish Sea, found an absence of genetic structuring data using microsatellite markers, and a study into *N. Nephrops* using allozyme electrophoresis was unable to detect genetic population structure at a broad scale across the European continent (Vendrami et al., 2017; McKeown et al., 2018; Stamatis et al., 2006). SNP analysis did however reveal fine scale structuring in *P. maximus*, and has been successfully used to discover a genetic cline in *H. gammarus*, (Jenkins et al., 2019). It would appear that the majority of marine species exhibiting a planktotrophic larval life-stage, have similar population structures, in Europe and the British Isles. A long-lived, larval life stage, whose movement is dictated largely by oceanic currents, appears to result in an absence of fine-scale genetic structuring in most cases, and we have evidence of broader scale structure occurring, which may be a common feature, but has yet to be recorded. It is highly conceivable that local hydrodynamics produce a barrier to the majority of larvae, in a system where larval mortality is extremely high. The species compared here all have similar life history traits in that they all exhibit hatching of eggs in the summer months and a planktotrophic life stage lasting several weeks. We can assume that species with larvae released at different times of year, and with a significantly shorter planktotrophic period, may be more likely to exhibit significantly more genetic population structure since the opportunity for dispersal is diminished.

Definitions of biological populations vary in the literature but include both demographic and reproductive interactions, and can be considered to be ecological or evolutionary definitions, respectively (Wapes and Gaggiotti, 2006). On a scale from complete isolation at one extreme, to panmixia at the other, there is a point on this scale where one can satisfyingly define that two (or more) populations have diverged sufficiently from panmixia to be considered separate populations. This subjective definition of population differentiation also leads to a lack of clarity upon the definition of evolutionarily significant units (McElhany et al., 2000). A useful definition relates to the number of effective migrants (N_em) moving between a pair of putatively distinct populations (Wright, 1931). Even a small number of migrants per generation ($N_em=1$) will reduce the random effects of genetic drift and cause increasing population cohesion through gene flow and decreasing values of F_{ST} (Mills and Allendorf, 1996; Wang, 2004). This very low number of effective migrants required to depress F_{ST} and increase population cohesiveness in the evolutionary sense, is in contrast to the relatively high

number of effective migrants required to maintain demographic homogeneity, which has been estimated to be around 10% of the population size (Hastings, 1993). With regards to the recovery of a stock experiencing high fishing pressure, the very low numbers of migrants ($N_em \approx 1$) required to cause depression of F_{ST} will have a negligible effect upon population recovery, however those sufficient to cause maintain demographic similarities (estimated at $N_em > 500$) are more likely to influence stock recovery. Based on the present calculations of pairwise F_{ST} between putative populations indicate that that the *H. gammarus* fisheries around the UK and Ireland are best considered to be a large, well-mixed population, at the western edge of their range. However, the translation of genetic/evolutionary isolation to demographic dispersal rates is complex (Spies et al., 2018). Migration rates are likely to be high in these *H. gammarus* fisheries (Rötzer and Haug, 2015), and both ecological and evolutionary homogeneity are likely in this case. Our findings are in keeping with other studies which have attempted to estimate both local population structure (Ellis et al., 2017; Watson et al., 2016) and broader structuring across the continent (Jenkins et al., 2019) using genetic markers. Management should be driven by legislation at an effective spatial scale. Our data suggests a well-mixed population around the UK, with high levels of gene flow between regions. Management should therefore be focused upon both a national scale as well as upon a regional scale, since recruitment and stock recovery between regions appear to be very closely linked with relatively high rates of migration.

4.6.2 Marker Choice for Population Genetics of *Homarus gammarus*.

The secondary goal of this study was to perform a comparison into the effectiveness of two types of marker for population genetics, specifically for research into the population dynamics of a species with a long-lived larval mode of dispersal. Both marker types tested gave almost identical results, both describing an absence of population structure. As such, we are unable report a preferential marker type for analysis of the population structure of *H. gammarus* with reference to statistical power, or suitability for analysis of a highly mixed population. Comparisons in the practicality and cost of each marker show that whilst microsatellites are likely to be a cheaper method of studying population genetics (considering reagents and consumables costs), methods using SNPs are much quicker allowing ecological questions to be answered more efficiently. Microsatellite analysis followed an established workflow, requiring a

significantly less intense workload but over a much more prolonged period. In total we analysed 383 DNA samples, using six multiplex primer mixes. These 2,298 unique PCR reactions were all synthesised and analysed manually, however this conservative estimate of the number of PCR reactions does not include any optimisations or repeats that were included in the study and as such we would estimate that the true number of unique reactions was closer to 4,000. In comparison, RAD-Seq and SNP analysis of 95 DNA samples required a more intense period of lab work over approximately three weeks, followed by a far less intense course of computer data analysis, characterised by running long computational jobs rather than scrutinising plots by eye. Given that both methods gave very similar outputs in the present study, we would suggest that in this case, RAD-Seq and SNP analysis would be the preferred method, primarily for the improved speed of statistical analysis and generation of results. An analysis of cost estimates associated with both microsatellite and SNP analysis of 96 DNA samples (Appendix 4: Table S1), shows a higher cost for a typical RAD-Seq type analysis, but does not include staff-time, which if included would render microsatellite analysis the more expensive option.

4.6.3 Appraisal of Methods

Using the methods described here we achieved very high rates of successful microsatellite genotyping, and generated high quality restriction site associated next-generation sequence data. Our multiplex PCR conditions for the 20 microsatellite markers were optimised over several iterations and consistently performed very well with no issues relating to spectral overlap, or stutter peaks which can cause issues with genotyping. The RAD-Seq library preparation performed well but generated less sequence data than anticipated, leading to the generation of relatively few SNP markers, in comparison to some SNP studies with a very high marker count (Berihulay et al., 2019; Cai et al., 2018). This resulted in a reduction in our statistical power to detect population structure. The threshold of only using markers which occurred in 40% of individuals is low for an SNP study of this type. This decision was made as a direct result of lack of sequence depth of coverage, relating to under-clustering on the flowcell. However, given that our results are consistent with those derived from the microsatellite component of this study, and other studies into the species, this does not appear to have negatively affected the validity of our results. Conservation decisions are often based upon low resolution data and any contribution that genetic data can make to species

conservation is always beneficial (Fox et al., 2018). There have now been several investigations into the *H. gammarus* population structure in several regions of the UK and Ireland, all of which have concluded that the population is panmictic, to a large degree (Watson et al., 2016; Ellis et al., 2017). Further increases in sample or genetic marker number are attractive. However, given our knowledge of the broad dispersal potential of *H. gammarus* (Rötzer and Haug, 2015; Werner, 2007), we argue that sufficient resolution has been achieved to provide conservation authorities with the required information to adequately manage the fisheries in the region.

Alternative, or complementary sampling strategies and analytical approaches may be required to provide the power, or diversity of information to confidently describe the population in the region, such as particle tracking or plankton tows. Many planktonic organisms have very little control over their movement in the oceans. Particle tracking approaches seek to model the oceanographic currents and track the movement of *in silico* larvae for the duration of the larval life stage in the target species and have been used alongside genetic data to identify broad migratory patterns, source and sink populations and identify physical barriers to gene flow (North et al., 2008; Stuckas et al., 2017). Plankton tows can also be employed to capture actively migrating plankton and perform a direct measure of dispersal distances, as opposed to metrics calculated by analysis of the settled adult population (Andrews et al., 2014). Provided a captured larvae can be assigned to a source population with high confidence using genetic markers, this provides a method of direct sampling of migrating larvae, and does not assume that all larvae which successfully migrate, survive and colonise (Salinas-de Leon et al., 2012). Finally, large sample size is often employed as a method to increase statistical power, and enable the detection of fine-scale, or weak genetic structure (Andrews et al., 2014; Holland et al., 2017). These are some examples of well-established alternative methods to complement genetic population analyses, and given the consensus of very low genetic differentiation in the *H. gammarus* populations of the UK and Ireland, some application of these methods may be required for any further enlightenment.

References

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B., Guerler, A., Hillman-Jackson, J., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., and Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, 46(W1):W537–W544.

Alberto, F. (2009). Msatallele_1.0: An R package to visualize the binning of microsatellite alleles. *Journal of Heredity*, 100(3):394–397.

Allendorf, F., Luikart, G., and Aitken, S. (2012). *Conservation and the Genetics of Populations*. Wiley-Blackwell, New Jersey, USA.

Andrews, K., Norton, E., Fernandez-Silva, I., Portner, E., and Goetze, E. (2014). Multilocus evidence for globally distributed cryptic species and distinct populations across ocean gyres in a mesopelagic copepod. *Molecular Ecology*, 23(22):5462–79.

André, C. and Knutsen, H. (2010). Development of twelve novel microsatellite loci in the European lobster (*Homarus gammarus*). *Conservation Genetics Resources*, 2(1):233–236.

Babbucci, M., Buccoli, S., Cau, A., Cannas, R., Goñi, R., Díaz, D., Marcato, S., Zane, L., and Patarnello, T. (2010). Population structure, demographic history, and selective processes: Contrasting evidences from mitochondrial and nuclear markers in the European spiny lobster *Palinurus elephas* (fabricius, 1787). *Molecular Phylogenetics and Evolution*, 56(3):1040–1050.

Balloux, F., Brunner, H., Lugon-Moulin, N., Hausser, J., and Goudet, J. (2000). Microsatellites can be misleading: an empirical and simulation study. *Evolution*, 54(4):1414–1422.

Barreiro, L., Laval, G., Quach, H., Patin, E., and Quintana-Murci, Q. (2008). Natural selection has driven population differentiation in modern humans. *Nature Genetics*, 40:340–345.

Benestan, L., Gosselin, T., Perrier, C., Sainte-Marie, B., Rochette, R., and Bernatchez, L. (2015). RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). *Molecular Ecology*, 24(13):3299–315.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188.

Berihulay, H., Li, Y., Liu, X., Gebreselassie, G., Islam, R., Liu, W., Jiang, L., and Ma, Y. (2019). Genetic diversity and population structure in multiple chinese goat populations using a SNP panel. *Animal Genetics*, 50(3):242–249.

Blacket, M., Robin, C., Good, R., Lee, S., and Miller, A. (2012). Universal primers for fluorescent labelling of PCR fragments - an efficient and cost effective approach to genotyping by fluorescence. *Molecular Ecology Resources*, 12(3):456–463.

Bolger, A., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Brookfield, K., Gray, T., and Hatchard, J. (2005). The concept of fisheries-dependent communities: A comparative analysis of four UK case studies: Shetland, Peterhead, North Shields and Lowestoft. *Fisheries Research*, 72(1):55–69.

Browne, R., Mercer, J., and Duncan, M. (2001). An historical overview of the Republic of Ireland's lobster (*Homarus gammarus* linnaeus) fishery, with reference to European and north American (*Homarus americanus* milne edwards) lobster landings. *Hydrobiologia*, 465(1-3):49–62.

Cai, S., Xu, S., Liu, L., Gao, T., and Zhou, Y. (2018). Development of genome-wide SNPs for population genetics and population assignment of *Sebastiscus marmoratus*. *Conservation Genetics Resources*, 10(3):575–578.

Casey, J., Jardim, E., and Martinsohn, J. (2016). The role of genetics in fisheries management under the e.u. common fisheries policy. *Journal of Fish Biology*, 89:2755–2767.

Catchen, J., Hohenlohe, P., Bassham, S., Amores, A., and Cresko, W. (2013). STACKS: an analysis tool set for population genomics. *Molecular Ecology*, 22(11):3124–40.

Chapuis, M. and Estoup, A. (2007). Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution*, 24(3):621–631.

Chapuis, M., Lecoq, Y., Michalakis, A., Loiseau, G., Sword, A., Piry, S., and Estoup, A. (2008). Do outbreaks affect genetic population structure? a worldwide survey in *Locusta migratoria*, a pest plagued by microsatellite null alleles. *Molecular Ecology*, 17(16):3640–3652.

Clark, C., Hughes, A., Greenwood, S., Jordan, C., and Sejrup, H. (2010). Pattern and timing of retreat of the last British-Irish ice sheet. *Quaternary Science Reviews*, pages 1–35.

Coates, B., Sumerford, D., Miller, N., and Kim, K. (2009). Comparative performance of single nucleotide polymorphism and microsatellite markers for population genetic analysis. *Journal of Heredity*, 100(5):556–564.

Cobb, J. and Castro, K. (2006). *Lobsters: Biology, Management, Aquaculture and Fisheries*. Blackwell Publishing.

Covarrubias-Pazaran, G., Diaz-Garcia, L., Schlautman, B., Salazar, W., and Zapala, J. (2016). Fragman: An R package for fragment analysis. *BMC Genetics*, 17(62):1–8.

Dempster, A., Laird, N., and Robin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Elliott, M. and Holden, J. (2017). UK Sea Fisheries Statistics 2017. *National Statistics*, (Marine Management Organisation).

Ellis, C., Hodgson, D., André, C., Sørdalen, T., Knutsen, H., and Griffiths, A. (2015). Genotype reconstruction of paternity in European lobsters (*Homarus gammarus*). *PLOS ONE*, 10(11):e0139585.

Ellis, C., Hodgson, D., Daniels, C., Collins, M., and Griffiths, A. (2017). Population genetic structure in european lobsters: implications for connectivity, diversity and hatchery stocking. *Marine Ecology Progress Series*, 563:123–137.

European Commission (2014). The Common Fisheries Policy [Accessed online: https://ec.europa.eu/fisheries/cfp$_en$], 08/10/2019.

Fischer, M., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F., Shimizu, K., Holderegger, R., and Widmer, A. (2017). Estimating genomic diversity and population differentiation -an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. *BMC Genomics*, 18.

Fox, G., Darolti, I., Hibbitt, J., Preziosi, R., Fitzpatrick, J., and Rowntree, J. (2018). Genetic assessment of *ex situ* populations to aid species conservation and maintain heterozygosity in non-model species. *Journal of Zoo and Aquarium Research*, 6(2):50–56.

Frankham, R., Ballou, J., and Briscoe, D. (2004). *A Primer of Conservation Genetics*. Cambridge University Press, Cambridge, USA.

Gerlach, G., Jueterbock, A., Kraemer, P., Depperman, J., and Harmand, P. (2010). Calculations of population differentiation based on GST and D: forget GST but not all of statistics! *Molecular Ecology*, 19(18):3845–3852.

Gillespie, J. (1998). *Population Genetics A Concise Guide*. The John Hopkins University Press, Baltimore, Maryland.

Gkagkavouzis, K., Karaiskou, N., Katopodi, T., Leonardos, I., Abatzopoulos, T., and Triantafyllidis, A. (2019). The genetic population structure and temporal genetic stability of gilthead sea bream (*Sparus aurata*) populations in the Aegean and Ionian Seas, using microsatellite DNA markers. *Journal of Fish Biology*, 94(4).

Goudet, J. (2005). hierfstat, a package for R to compute and test hierarchical F statistics. *Molecular Ecology Notes*, 5:184–186.

Griffiths, S., Fox, G., Briggs, P., Donaldson, I., Hood, S., Richardson, P., Leaver, G., Truelove, N., and Preziosi, R. (2016). A Galaxy-based bioinformatics pipeline for optimised, stramlined microsatellite development from Illimina next-generation sequencing data. *Conservation Genetics Resources*, 8:481–486.

Gärke, C., Ytournel, F., Bed'hom, B., I., G., Lathrop, M., Weigend, S., and Simianer, H. (2012). Comparison of SNPs and microsatellites for assessing the genetic structure of chicken populations. *Animal Genetics*, 43(4):419–28.

Hartl, D. and Clark, G. (1997). *Principles of Population Genetics*. Sinauer Associates, Inc. Publishers.

Hastings, A. (1993). Complex interactions between dispersal and dynamics: Lessons from coupled logistic equations. *Ecology*, 74:1362–1372.

Hastings, A. and Botsford, Louis, W. (2006). Persistence of spatial populations depends on returning home. *PNAS*, 103(15):6067–6072.

Hedrick, P. (1999). Perspective: Highly variable loci and their interpretation in evolution and conservation. *Evolution*, 53(2):313–318.

Helyar, S., Hemmer-Hansen, J., Bekkevold, D., Taylor, M., Ogden, R., Limborg, M., Cariani, A., Maes, G., Diopere, E., Carvalho, G., and Nielsen, E. (2011). Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources*, 11(s1):123–136.

Hill, A., Brown, J., and Fernand, L. (1997). The Summer Gyre in the Western Irish Sea: Shelp Sea Paradigms and Management Implications. *Estuarine, Coastal and Shelf Science*, 44(A):83–95.

Hill, W. (2009). Estimation of effective population size from data on linkage disequiilibrium. *Genetics Research*, 38(3):209–216.

Hill, W. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38:226–231.

Holland, L., Jenkins, T., and Stevens, J. (2017). Contrasting patterns of population structure and gene flow facilitate exploration of connectivity in two widely distributed temperate octocorals. *Heredity*, 119(1):35–48.

Holthuis, L. (1991). Key to species *Homarus americanus*, *H. capensis* and *H. gammarus*. *FAO Species Catalogue*, 13: Marine Lobsters of the World.

Hoorn, C., Wesselingh, F., ter Steege, H., Bermudez, M., Mora, A., Sevink, J., Sanmartin, I., Sanchez-Meseguer, A., Anderson, C., Figueiredo, J., Jaramillo, C., Riff, D., Negri, F., Hooghiemstra, H., Stadler, T., Sarkinen, T., and Antonelli, A. (2010). Amazonia through time: Andean uplist, climate change, landscape evolution and biofiversity. *Science*, 330(6006):927–931.

Ingebrigt, U., Benavente, G., and Browne, R. (2005). A regional development strategy for stock enhancement of clawed lobsters (*Homarus gammarus*): Development of juvenile

lobster production methodologies. Norwegian Institute for Nature Research(NINA Report).

Jeffries, D., Copp, G., Handley, L., K.H., O., C.D., S., and B., H. (2016). Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, (*Carassius carassius, L.*). *Molecular Ecology*, 25(13):2997–3018.

Jenkins, T., Ellis, C., Triantafyllidis, A., and Stevens, J. (2019). Single nucleotide polymorphisms reveal a genetic cline across the north-east Atlantic and enable powerful population assignment in the European lobster. *Evolutionary Applications*, 12(10).

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24:1403–1405.

Jombart, T. and Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27(21).

Jost, L. (2009). D vs. GST: Response to Heller and Siegismund (2009) and Ryman and Leimar (2009). *Molecular Ecology*, 18(10).

King, D., Ferguson, A., and Moffett, I. (1987). Aspects of the population genetics of herring, *Clupea harengus*, around the British Isles and in the Baltic Sea. *Fisheries Research*, 6:561–568.

Kleiven, A., Moland, E., Olsen, E., and Knutsen, J. (2018). *Integrated Coastal Zone Management*, chapter Lobster Reserves in Coastal Skagerrak – An Integrated Analysis of the Implementation Process. Springer Vieweg.

Lemopoulos, A., Prokkola, J., Uusi-Heikkilä, S., Vasemägi, A., Huusko, A., Hyvärinen, P., Koljonen, M., Koskiniemi, J., and Vainikka, A. (2019). Comparing RADseq and microsatellites for estimating genetic diversity and relatedness — Implications for brown trout conservation. *Ecology and Evolution*, 9(4):2106–2120.

Luikart, G. and Cornuet, J. (1999). Estimating the Effective Numbers of Breeders From Heterozygote Excess in Progeny. *Genetics*, 151(3):1211–1216.

Luikart, G., England, P., Tallmon, D., Jordan, S., and Taberlet, P. (2003). The power and promise of poppulation genomics: from genoptyping to genome typing. *Nature Reviews Genetics*, 4:981–994.

McElhany, P., Ruckelshaus, M., Ford, M., Wainwright, T., and Bjorkstedt, E. (2000). Viable salmonid populations and the recovery of evolutionarily significant units. Technical report, Northwest Fisheries Science center, U.S. Department of Commerce.

McKeown, N., Watson, H., Coscia, I., Wootton, E., and Ironsside, J. (2018). Genetic variation in Irish Sea brown crab (*Cancer pagurus* l.): implications for local and regional management. *Journal of the Marine Biological Association of the United Kingdom*, 99(4):879–886.

McPherson, A., O'Reilly, P., and Taggart, C. (2011). Genetic differentiation, temporal stability, and the absence of isolation by distance among Atlantic herring populations. *Transactions of the American Fisheries Society*, 133(2):434–446.

Mesak, F., Tatarenkov, A., Earley, R., and Avise, J. (2014). Hundreds of SNPs vs. dozens of SSRs: which dataset better characterizes natural clonal lineages in a self-fertilizing fish? *Frontiers in Ecology and Evolution*, 12.

Mills, L. and Allendorf, F. (1996). The one-migrant-per-generation rule in conservation and management. *Conservation Biology*, 6:1509–1518.

Morin, P., Archer, F., Pease, V., Hancock-Hanser, B., Robertson, K., Huebringer, R., Martien, K., Bickham, J., George, J., Postma, L., and Taylor, B. (2012). An empirical comparison of SNPs and microsatellites for population structure, assignment, and demographic analyses of bowhead whale populations. *Endangered Species Research*, 19:129–147.

Ménesguen, A. and Gohin, F. (2006). Observation and modelling of natural retention structures in the English Channel. *Journal of Marine Systems*, 63(3-4):244–256.

North, E., Schlag, Z., Hood, R., Li, M., Zhong, L., Gross, T., and Kennedy, V. (2008). Vertical swimming behaviour influecnes the dispersal of simularted oyster larvae in a coupled particle-tracking and hydrodynamics model of Chesapeake Bay. *Marine Ecology Progress Series*, 359:99–115.

Pandian, T. (1970). Ecophysiological studies on the developing eggs and embryoes of the European lobster *Homarus gammarus*. *Marine Biology*, 5:154–167.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, 26:419–420.

Paris, J., Stevens, J., and Catchen, J. (2017). Lost in parameter space: a road map for STACKS. *Methods in Ecology and Evolution*, 8(10).

Pasqualotto, A., Denning, D., and Anderson, M. (2007). A cautionary tale: lack of consistency in allele sizes between two laboratories for a published multilocus microsatellite typing system. *Journal of Clinical Microbiology*, 45(2):522–528.

Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotypes data. *Genetics*, 155(2):945–959.

Pudovkin, A., Zaykin, D., and Hedgecock, D. (1996). On the Potential for Eestimating the Effective Number of Breeders from Heterozygote-Excess in Progeny. *Genetics*, 144(1):383–387.

Rico, C., Cuesta, J., Drake, P., Macpherson, E., Bernatchez, L., and Marie, A. (2017). Null alleles are ubiquitous at microsatellite loci in the wedge clam (*Donax trunculus*). *PeerJ*, 5(e3188).

Roques, S., Chancerel, E., Boury, C., Pierre, M., and Acolas, M. (2019). From microsatellites to single nucleotide polymorphisms for the genetic monitoring of a critically endangered sturgeon. *Ecology and Evolution*, 9(12).

Rousset, F. (2008). genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, 8(1):103–106.

Rötzer, M. and Haug, J. (2015). Larval development of the European lobster and how small heterochronic shifts lead to a more pronounced metamorphosis. *Internation Journal of Zoology*, 2015(Article ID 345172).

Salinas-de Leon, P., Jones, T., and Bell, J. (2012). Successful determination of larval dispersal distances and subsequent settlement for long-lived pelagic larvae. *PLOS ONE*, 7(3):e32788.

Saunders, D., Hobbs, R., and Margules, C. (1990). Biological consequences of ecosystem fragmentation: A review. *Conservation Biology*, 5(1):18–32.

Schmalenback, I. and Buchholz, F. (2010). Vertical positioning and swimming performance of lobster larvae *Homarus gamarus* in an artificial water column at Helgoland, North Sea. *Marine Biology Research*, 6(1):89–99.

Sea-Fisheries Protection Authority (2017). Annual Landings Statistics; 2017 Landings [Accessed online: http://www.sfpa.ie/sea-fisheries-conservation/fisheries-statistics-and-quota-uptake/annual-landing-statistics/2017-landings], 17/01/2019.

Selkoe, K. and Toonen, R. (2006). Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters*, 9(5):615–29.

Smith, B., McCormack, J.E. Cuervo, A., Hickerson, M., Aleixo, A., Cadena, C., Pérez-Emán, J., Burney, C., Xie, X., Harvey, M., Faircloth, B., Glenn, T., Derryberry, E., Prejean, J., Fields, S., and Brumfield, R. (2014). The drivers of tropical speciation. *Nature*, 515:406–409.

Souza, T., Luna, L., Araripe, J., Melo, M., Silva, W., Schneider, H., Sampaio, I., and Rego, P. (2019). Characterization of the genetic diversity and population structure of the manakin genus (*Antilophia*) through the development and analysis of microsatellite markers. *Journal of Ornithology*, 160:825–830.

Spies, I., Hauser, L., Jorde, P., Knutsen, H., Punt, A., Rogers, L., and Stenseth, N. (2018). Inferring genetic connectivity in real populations exemplified by coastal and oceanic Atlantic cod. *Proceedings of the Natural Academy of Sciences*, 115(19):4945–4950.

Stamatis, C., Triantafyllidis, A., Moutou, K., and Mamuris, Z. (2006). Allozymic variation in Northeast Atlantic and Mediterranean populations of Norway lobster, *Nephrops norvegicus*. *ICES Journal of Marine Science*, 63(5):875–882.

Stuckas, H., Knöbel, L., Schade, H., Breusing, C., Hinrichsen, H., Bartel, M., Langguth, K., and Melzner, F. (2017). Combining hydrodynamic modelling with genetics: can passive larval drift shape the genetic structure of Baltic *Mytilus* populations? *Molecular Ecology*, 26(10):2765–2782.

Szpiech, Z., Jakobsson, M., and Rosenberg, N. (2008). ADZE: a rarefaction approach for counting alleles private combinations of populations. *Bioinformatics*, 24(21):2498–2504.

Taylor, A. (1995). North-South shifts of the Gulf Stream and their climatic connection with the abundance of zooplankton in the UK and its surrounding seas. *ICES Journal of Marine Science*, 52(3-4):711–721.

Team, R. C. (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna(Austria).

Truelove, N., Box, S., Behringer, D. J., Butler, M., and Preziosi, R. (2013). Genetic population structure of Caribbean spiny lobster. *Proceedings of the 66th Gulf and Caribbean Fisheries Institute*, pages 452–453.

Vendrami, D., Telesca, L., Weigand, H., Weiss, M., Fawcett, K., Lehman, K., Clark, M., Leese, F., McMinn, C., Moore, H., and Hoffman, J. (2017). RAD sequencing resolves fine-scale population structure in a benthic invertebrate: implications for understanding phenotypic plasticity. *Royal Society Open Science*, 4(2):160548.

Vieira, M., Santini, L., Diniz, A., and Munhoz, C. F. (2016). Microsatellite markers: what they mean and why they are so useful. *Genetics and Molecular Biology*, 39(3):312–328.

Vignal, A., Milan, D., Sancristobal, M., and Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution*, 34(3):275–305.

Wagner, D., Baris, T., Dayan, D., Oleksiak, M., and Crawford, D. (2017). Fine-scale genetic structure due to adaptive divergence among microhabitats. *Heredity*, 118(6):594–604.

Wang, J. (2004). Application of the one-migrant-per-generation rule to conservation and management. *Conservation Biology*, 18(2):332–343.

Wang, J. (2005). Estimation of effective population sizes from data on genetic markers. *Philos Trans R Soc Lond B Biol Sci*, 360(1459):1395–1409.

Wang, J. (2009). Deviation from Hardy-Weinberg proportions in finite populations. *Genetics Research*, 68(3):249–257.

Wapes, R. and Gaggiotti, O. (2006). What is a population? an empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, 15(6):1419–39.

Watson, H. V., McKeown, N. J., Coscia, I., Wootton, E., and Ironside, J. I. (2016). Population genetic structure of the European lobster (*Homarus gammarus*) in the Irish sea and implications for the effectiveness of the first British marine protected area. *Fisheries Research*, 183:287–293.

Werner, F. (2007). Population connectivity in marine systems: an overview. *Oceanography*, 20(2007):14–21.

Willing, E., Dreyer, C., and van Oosterhout, C. (2012). Estimates of genetic differentiation measured by F_{ST} do not necessarily require large sample sizes when using many SNP markers. *PLOS ONE*, 7(8):e42649.

Willoughby, J., Harder, A., Tennessen, J., Scribner, K., and Christie, M. (2018). Rapid genetic adaptation to a novel environment despite a genome-wide reduction in genetic diversity. *Molecular Ecology*, 27(20):4041–4051.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16:97–159.

Wu, L., Wen, C., Qin, J., Yin, H., Tu, Q., Van Nostrand, J., Yuan, T., Yuan, M., Deng, Y., and Zhou, J. (2015). Phasing amplicon sequencing on Illumina Miseq for robust environmental microbial community analysis. *BMC Microbiology*, 15(125).

Young, A., Boyle, T., and Brown, T. (1996). The population genetic consequences of habitat fragmentation for plants. *Trends Ecol Evol*, 11(10):413–8.

Zheng, X., Pierce, G., Reid, D., and Jolliffe, I. (2002). Does the North Atlantic current affect spatial distribution of whiting? Testing environmental hypotheses using statistical and GIS techniques. *ICES Journal of Marine Science*, 59(2):239–253.

4.7 Acknowledgements

Sample collection from sites around the UK and Ireland was made possible by a great many people whom were extremely generous with their time. In no particular order, our gratitude goes to: Ms. Jane McMinn of Firth of Forth Lobster Hatchery (JM), Dr. Charlie Ellis of The National Lobster Hatchery (CE), Mr. Jason Sparks (JS) and Mr. Jonathan Haines (JH) of North Western Inshore Fisheries and Conservation Authority, Ms. Sally Stewart Moore of Northumberland Inshore Fisheries and Conservation Authority (SSM), Dr. Sarah Perry of The Wildlife Trust of South and West Wales (SP), Dr. Oliver Tully of The Marine Institute (Ireland) (OT) and Sean Faulkner of Faulkner Fisheries (SF).

4.8 Conflicts of Interest

The authors declare that they have no conflict of interest.

4.9 Tables and Figures

Figure 4.1

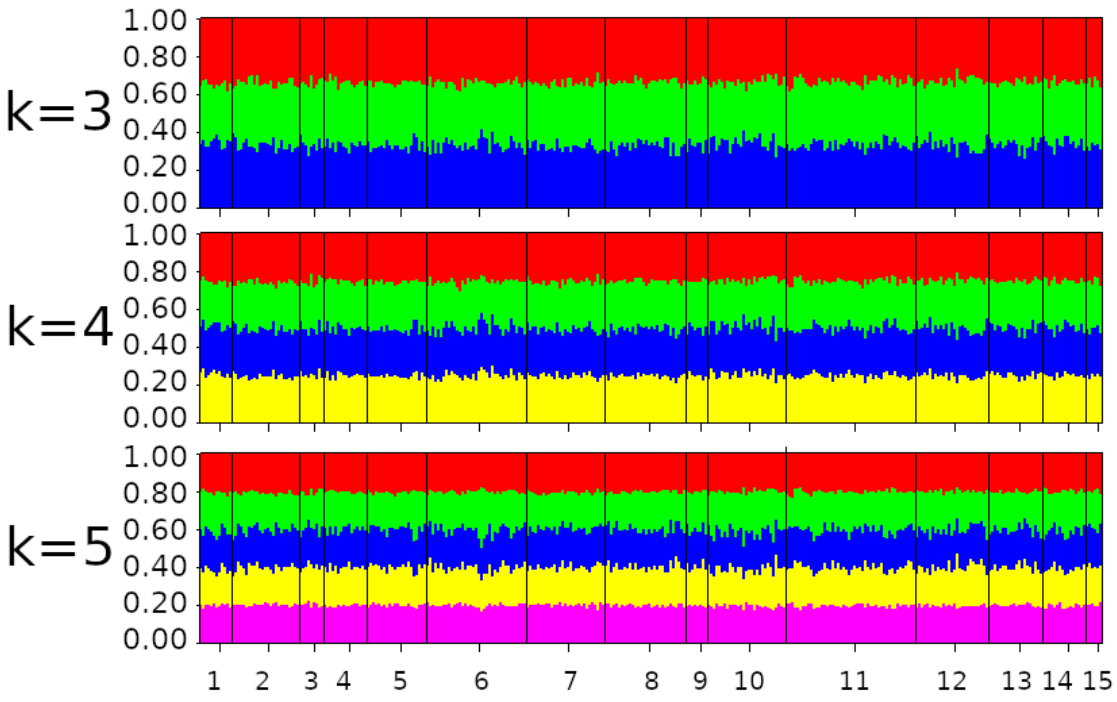


Figure 4.2

Table 4.1

Sampling site, sampling site ID, broader site ID (NE=North East, SW=South West, IS=Irish Sea, AI=Atlantic Ireland), approximate latitude and longitude, N (µSAT): number of samples successfully genotyped with microsatellite markers and used in analysis, N (SNP): number of samples analysed using SNP markers, year of sampling, Sampling coordinator (see acknowledgements).

Table 4.1

Map ID	Site	ID	Broad Site	Latitude	Longitude	N (µSAT)	N (SNP)	Year	Sampling Coordinator
1	County Clare	CCL	IS	52.895294	-9.532182	12	12	2015	OT
2	Donegal	DON	AI	54.555871	-8.377615	25	12	2015	OT
3	Waterford	WAT	IS	52.169763	-6.936047	9	0	2015	OT
4	Wexford	WEX	IS	52.316916	-6.325791	16	0	2015	OT
5	Mid-Irish Sea	MIS	IS	53.332399	-5.302501	25	12	2015	OT
6	St. Bees	BEE	IS	54.514077	-3.651970	37	12	2016	JS & JH
7	Isles of Scilly	IOS	SW	49.935254	-6.321319	30	0	Pre-2014	CE
8	Portreath	POR	SW	50.286217	-5.308941	30	12	Pre-2014	CE
9	Boscastle	BOS	SW	50.718634	-4.741606	8	12	Pre-2014	CE
10	Lizard Point	LIZ	SW	49.956634	-5.205915	29	0	Pre-2014	CE
11	Firth of Forth	FOF	NE	56.130284	-2.763474	66	0	2014-2017	JM
12	Seahouses	SEA	NE	55.585092	-1.650097	33	0	2015, 2016	CE
13	Craster	CRA	NE	55.474056	-1.590104	28	0	2015, 2016	SSM
14	Amble	AMB	NE	55.337737	-1.579213	24	12	2015, 2016	SSM
15	North Shields	NSH	NE	55.009177	-1.415975	11	11	2015, 2016	SSM

Table 4.2

Locus	Primer Sequence (5′- 3′)	N	A	T_A°C	Motif	Range (bp)
GFHG01	F: ATTACTGCTGGGTAGACAGAGG	277	16	60	ATAC	451-509
	R: GTTAAGGAGGAGGTAAAAGGTAGG					
GFHG09	F: ACCTCAGTCTAGATCATACACTGG	308	13	60	ATAC	167-204
	R: GTGTGTGACTAGCAGATAGATGC					
GFHG11	F: AGGAGTAAGACATCTCCATACACC	290	8	60	ATAC	410-430
	R: TTCTGATCCCAGCAATACTCC					
GFHG13	F: GTCCTCGTGTACAATAGTGGG	277	22	60	ATCT	233-307
	R: GAGATAATGTTGAGGAAGAGGG					
GFHG16	F: GTGTAGGTGACGTATGACTGTCG	279	18	60	ATGT	422-472
	R: AGAGAAGTAGACAGATAGGATGGC					
GFHG30	F: ACACTAAACGCTACCACACTAGACG	296	6	60	ATCT	488-512
	R: ATGACTTTATTACGCGGGACC					

Table 4.3

Multiplex ID	Marker	Tail	Dye	T_A°C
Mplex1	HGC129	M13_A	TAMRA	58
	HGC103	M13_B	PET	58
	HGB6	Blkt_C	FAM	58
	HGD111	Blkt_C	FAM	58
Mplex2	HGC118	M13_A	TAMRA	58
	HGA8	Blkt_C	FAM	58
	HGC131b	M13_B	PET	58
	HGD106	M13_B	PET	58
Mplex3	HGB4	Blkt_C	FAM	58
	HGC6	Blkt_C	FAM	58
	HGC120	M13_B	PET	58
Mplex4	HGD110	Blkt_C	FAM	58
	HGD117	M13_A	TAMRA	58
	HGD129	M13_B	PET	58
Mplex5	GFHG13	M13_A	TAMRA	60
	GFHG09	Blkt_C	FAM	60
	GFHG01	Blkt_C	FAM	60
Mplex6	GFHG16	M13_A	TAMRA	60
	GFHG11	M13_B	PET	60
	GFHG30	Blkt_C	FAM	60

Table 4.4

Locus	Mean NA	H_{OBS}	H_{EXP}	F_{ST}
HGC120	0.008	0.858	0.861	0.002
HGB4	0.013	0.662	0.640	0.003
HGD106	0.010	0.702	0.709	0.003
HGC6	0.002	0.347	0.338	0.002
HGA8	0.074	0.655	0.803	0.000
HGC103	0.021	0.690	0.702	0.001
HGC118	0.016	0.614	0.612	0.000
HGC131b	0.007	0.826	0.820	0.000
HGC129	0.046	0.680	0.756	0.001
HGD111	0.033	0.535	0.567	0.002
HGB6	0.015	0.736	0.711	0.006
HGD110	0.012	0.800	0.806	0.003
HGD117	0.039	0.497	0.574	0.000
HGD129	0.018	0.598	0.615	0.002
GFHG01	0.022	0.590	0.621	0.002
GFHG09	0.035	0.684	0.708	0.001
GFHG13	0.030	0.771	0.883	0.001
GFHG30	0.017	0.390	0.422	0.001

Table 4.5

Sampling location, N: Sample number (after removal of failed samples), AR *(SE)*: average rarefied allelic richness *(SE)*, P_A: absolute number of private alleles, PAR *(SE)*: average rarefied private allelic richness *(SE)*, H_{OBS}: observed heterozygosity, H_{EXP}: observed heterozygosity, F_{IT} *(SE)*: average inbreeding coefficient *(SE)*.

Table 4.5

Site	N	AR (SE)	P_A	PAR (SE)	H_{OBS}	H_{EXP}	F_{IT} (SE)
Amble	15	2.188	1	0.153	0.643	0.653	0.168
		(0.073)		*(0.023)*			*(0.023)*
Boscastle	8	2.195	2	0.119	0.636	0.631	0.163
		(0.086)		*(0.037)*			*(0.016)*
County Clare	11	2.141	0	0.080	0.626	0.619	0.188
		(0.088)		*(0.010)*			*(0.034)*
Craster	18	2.205	4	0.154	0.659	0.661	0.161
		(0.082)		*(0.032)*			*(0.016)*
Donegal	25	2.217	3	0.140	0.662	0.677	0.154
		(0.069)		*(0.033)*			*(0.009)*
Firth of Forth	42	2.185	5	0.154	0.638	0.664	0.176
		(0.079)		*(0.019)*			*(0.010)*
Isles of Scilly	28	2.223	2	0.130	0.684	0.679	0.153
		(0.077)		*(0.024)*			*(0.012)*
Lizard Point	29	2.197	2	0.149	0.638	0.668	0.166
		(0.076)		*(0.022)*			*(0.011)*
Mid-Irish Sea	19	2.229	2	0.162	0.696	0.672	0.146
		(0.086)		*(0.027)*			*(0.015)*
North Shields	5	2.273	0	0.190	0.706	0.634	0.134
		(0.091)		*(0.045)*			*(0.033)*
Portreath	30	2.201	6	0.169	0.659	0.666	0.159
		(0.088)		*(0.027)*			*(0.011)*
Seahouses	25	2.240	3	0.155	0.671	0.687	0.164
		(0.066)		*(0.029)*			*(0.013)*
St. Bees	37	2.206	1	0.138	0.659	0.675	0.166
		(0.073)		*(0.026)*			*(0.011)*
Waterford	9	2.216	1	0.156	0.655	0.653	0.149
		(0.074)		*(0.018)*			*(0.016)*
Wexford	16	2.265	2	0.144	0.695	0.695	0.151
		(0.058)		*(0.038)*			*(0.015)*

Table 4.6

Probability of departure from Hardy-Weinberg equilibrium for each *H. gammarus* population and locus. Significant P-values after "B-Y" false discovery rate correction highlighted in bold.

Table 4.6

Pop \ Locus	HGC120	HGB4	HGD106	HGC6	HGA8	HGC103	HGC118	HGC131b	HGC129	HGD111	HGB6	HGD110	GFHG01	GFHG09	GFHG11	GFHG13	GFHG16	GFHG30
AMB	1.000	1.000	1.000	0.109	1.000	1.000	1.000	1.000	1.000	1.000	0.650	1.000	1.000	1.000	1.000	1.000	1.000	1.000
BOS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	NA	1.000	1.000	NA	1.000
CCL	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CRA	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.568	1.000	0.471	1.000	1.000	1.000	1.000	1.000	0.568	1.000	NA
DON	1.000	1.000	1.000	1.000	0.1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.195	**0.000**	1.000	1.000	1.000	1.000
FOF	1.000	**0.002**	**0.000**	1.000	1.000	1.000	**0.023**	**0.000**	0.207	**0.000**	**0.005**	**0.005**	1.000	**0.000**	1.000	1.000	1.000	1.000
IOS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.000**	0.109	1.000	1.000	1.000	1.000	1.000	0.334	1.000	1.000
LIZ	1.000	1.000	1.000	1.000	0.129	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.002**	1.000	1.000	1.000	1.000	1.000
MIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
NSH	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	NA
POR	1.000	1.000	1.000	1.000	1.000	**0.033**	1.000	1.000	1.000	1.000	1.000	1.000	0.328	1.000	0.354	1.000	1.000	1.000
SEA	1.000	1.000	0.139	1.000	0.102	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.201	1.000	1.000
BEE	1.000	1.000	1.000	1.000	1.000	1.000	**0.000**	1.000	**0.000**	1.000	1.000	1.000	0.863	1.000	1.000	0.210	**0.014**	1.000
WAT	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.423
WEX	1.000	1.000	0.119	0.421	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.421	1.000	1.000	1.000	1.000	1.000

Table 4.7

Locus	Population	Estimated Null Allele Frequency
HGA8	Wexford	0.10009
HGC103	Boscastle	0.10032
HGD129	Donegal	0.10325
GFHG13	Lizard Point	0.10365
GFHG11	Amble	0.10437
GFHG11	North Shields	0.10539
GFHG13	Boscastle	0.10539
GFHG30	Isles of Scilly	0.10711
HGC131b	Firth of Forth	0.11280
HGC103	Donegal	0.11293
HGA8	Boscastle	0.12139
HGD117	St. Bees	0.12385
HGD117	Craster	0.12862
HGC129	Craster	0.13354
HGB6	County Clare	0.13645
HGA8	Mid Irish Sea	0.13649
HGD117	Donegal	0.13707
HGA8	County Clare	0.15234
HGA8	Portreath	0.15321
GFHF09	Boscastle	0.16625
HGD111	North Shields	0.29029

Table 4.8: Pairwise F_{ST} and D for each pair of sampled *H. gammarus* sites. F_{ST} (above diagonal) values 0.05-0.15 highlighted in bold typeface. Hartl & Clark (1997) define a F_{ST} value in the range 0.05-0.15 as describing moderate genetic differentiation. Jost's D (below diagonal, shaded gray) showed no significant values.

Table 4.8

	AMB	BOS	CCL	CRA	DON	FOF	IOS	LIZ	MIS	NSH	POR	SEA	BEE	WAT	WEX
AMB		0.037	0.016	0.017	0.015	0.005	0.014	0.011	0.012	0.028	0.010	0.009	0.011	0.023	0.021
BOS	0.080		**0.053**	0.038	0.032	0.019	0.025	0.028	0.036	**0.098**	0.023	0.029	0.023	**0.065**	0.047
CCL	-0.020	0.080		0.017	0.020	0.010	0.011	0.009	0.014	0.031	0.010	0.012	0.011	0.028	0.027
CRA	0.001	0.075	-0.004		0.019	0.011	0.015	0.014	0.016	0.021	0.009	0.012	0.012	0.020	0.022
DON	0.006	0.101	0.019	0.031		0.013	0.014	0.015	0.013	0.024	0.013	0.011	0.015	0.019	0.021
FOF	-0.009	0.055	-0.002	0.007	0.017		0.013	0.010	0.007	0.016	0.007	0.009	0.008	0.014	0.018
IOS	0.011	0.086	-0.012	0.018	0.014	0.021		0.011	0.010	0.018	0.010	0.011	0.011	0.012	0.016
LIZ	-0.008	0.089	-0.020	0.014	0.020	0.008	0.006		0.009	0.020	0.011	0.011	0.012	0.015	0.015
MIS	-0.002	0.088	-0.011	0.004	0.016	0.000	-0.002	-0.002		0.020	0.010	0.011	0.007	0.016	0.014
NSH	-0.007	0.105	-0.036	-0.030	0.038	0.013	-0.010	0.022	-0.013		0.017	0.017	0.013	0.035	0.023
POR	-0.001	0.056	-0.011	-0.006	0.020	0.000	0.008	0.007	0.001	-0.002		0.007	0.008	0.016	0.015
SEA	-0.012	0.079	-0.018	-0.002	0.005	-0.002	0.001	0.004	-0.005	-0.019	-0.010		0.008	0.016	0.016
BEE	0.005	0.094	-0.002	0.013	0.021	0.005	0.008	0.013	-0.006	-0.006	0.004	-0.002		0.013	0.015
WAT	0.009	0.072	-0.008	0.002	0.021	0.008	-0.011	0.006	-0.001	-0.024	0.016	0.002	0.006		0.024
WEX	0.013	0.109	0.022	0.021	0.033	0.028	0.015	0.016	0.005	-0.032	0.017	0.011	0.020	0.018	

Table 4.9

Microsatellite Analysis					SNP Analysis				
	NE	SW	AI	IS		NE	SW	AI	IS
NE		0.005	0.007	0.002	NE		0.001	-1.466	-2.153
SW	0.001		0.007	0.002	SW	0.001		-1.933	-2.153
AI	-0.000	0.003		0.007	AI	-0.000	0.003		-1.951
IS	0.001	0.000	0.002		IS	0.001	0.000	0.002	

Table 4.10

	Microsatellite Analysis						SNP Analysis						
Site	N	AR *(SE)*	P_A	PAR *(SE)*	H_{OBS}	H_{EXP}	N	Reads	AR *(SE)*	P_A	PAR *(SE)*	H_{OBS}	H_{EXP}
North East	105	2.162 *(0.074)*	14	0.644 *(0.132)*	0.653	0.681	22	1,445,591	30.410 *(0.406)*	128	41.182 *(0.094)*	0.160	0.582
South West	95	2.151 *(0.076)*	9	0.560 *(0.137)*	0.658	0.680	21	1,394,481	15.048 *(0.513)*	37	5.381 *(0.029)*	0.111	0.780
Atlantic Ireland	55	2.158 *(0.074)*	3	0.385 *(0.101)*	0.667	0.680	23	3,017,422	65.826 *(0.289)*	931	413.913 *(0.272)*	0.178	0.242
Irish Sea	62	2.183 *(0.063)*	2	0.427 *(0.106)*	0.670	0.693	23	2,225,005	15.957 *(0.550)*	66	11.957 *(0.054)*	0.116	0.755

Chapter 5

An Analysis of Bias in Plant Barcoding Markers Using Next-Generation Sequencing.

5.1 Are all barcoding markers equal? Comparisons between plant metabarcoding markers for honey analysis.

Graeme Fox,[1] Richard F. Preziosi,[1] Loreto Ros,[2] Joshua Sammy,[1] Jennifer K. Rowntree[1] and Latha R. Vellaniparambil[1]

[1]Ecology and Environment Research Centre, Department of Natural Sciences, Manchester Metropolitan University, Chester Street, Manchester, United Kingdom, M1 5GD

[2]Faculty of Life Sciences, The University of Manchester, Oxford Road, Manchester M13 9PL

Keywords: metabarcoding, next-generation sequencing, palynology, pollen, bias, marker choice

Author contributions: GF, LRV, RFP and JKR conceived and designed the project; LRV collected the samples; LRV, YG LR and JS performed the lab work; GF performed the computer programming; GF performed the data analysis; GF wrote the chapter.

5.2 Abstract

Understanding the diet of pollinating insects is critical to manage the ecological threats facing them. Metabarcoding of plant DNA from honey can provide information on the community of forage plants utilised over a period of time. While an effective method for acquiring community level data in many ecosystems, the list of factors affecting community analysis by metabarcoding is long and must be properly understood for confidence in data interpretation. Here we investigate the relative biases in two plant metabarcoding markers (*rbcL* and *ITS2*), using them in parallel to characterise the pollen in honey sampled from hives in Greater Manchester. We use a bespoke bioinformatics pathway to analyse data, assigning species taxonomy to as many reads as possible. Comparison of assigned taxonomies to a custom database of species plausibly detected in the UK allows analysis of the false-positive rate of species detection, which we found to be extremely high. We observe high levels of variation between descriptions of the honey derived plant DNA from *rbcL* and *ITS2* with communities described by each marker significantly different from one another. We conclude that each marker individually does not allow one to make confident species level assignments in many instances. By using two markers in parallel, we are able to increase the confidence with which we can assign species, and increase the scope of taxa detection through the use of divergent markers. We highlight the caution which much be exercised when performing metabarcoding of plant samples, but our results are applicable to any metabarcoding experiment.

5.3 Introduction

The highly publicised decline in bees and other pollinating insects is an emotive and politically active subject, the importance of which has been brought into sharp focus with policy makers and the general public in recent decades (Department for Environment Food and Rural Affairs, 2014). Pollinating insects play a crucial role in the production of food crops (Hung et al., 2018; Gallai et al., 2009), and habitat loss, widespread pesticide use and adverse climatic conditions are known to be among the most powerful drivers of their decline (Potts et al., 2010; Pound et al., 2017; Schultz and Dlugosch, 1999; Goulson et al., 2005). Pollution from the use of agro-chemicals, including pesticides and insecticides, and in particular the systemic neonicotinoids, have well documented links to pollinator declines. The broad-spectrum effectiveness and persistence of neonicotinoids in soils, and

their water solubility have lead to them also being responsible for the destruction of many beneficial arthropods, and bioaccumulation also leads to mortality in birds and mammals further up the food chain (Maini et al., 2010; Goulson, 2013; Godfray et al., 2014). The widespread adoption of neonicotinoids in the agricultural industry quickly lead to the discovery that toxic residues of the active ingredient were being carried into bee colonies through the ingestion of pollen and nectar from treated crops (Sgolastra et al., 2020). Affected colonies experienced foraging bees losing their ability to return to the hive, and overall reductions in colony growth and production (Henry et al., 2012; Gill et al., 2012). Analysis of the pollen and nectar diet of foraging bees enables analysis of the importance of agricultural crops to the diet of hives, and their risk of carrying toxins back to the colony.

As intensification of farming processes continues, and globally more habitat is lost to agriculture (Kienast et al., 2019), bees and pollinators are under threat by the resulting reductions of quality, quantity and diversity of floral resources (Bloom et al., 2019; Goulson et al., 2015). Agricultural intensification has been closely linked with loss of diversity of plant species, and a reduction in abundance and diversity of bees, in particular where production is focused upon animal husbandry (Féon et al., 2010). Quality nutrition, through collecting an abundance of diverse pollen, the main source of protein for honey bees, is vital for healthy colonies to persist and thrive (Topitzhofer et al., 2019). Increasing our understanding of the interactions between bees and the flowering plants they visit is vital to inform their conservation, and by using the isolation and identification of pollen extracted from honey, we are able to analyse their diets in a relatively non-invasive manner. The development of pollen analysis methods have been particularly important to the study of pollinator ecology as these methods allow researchers valuable insight into the forage characteristics of a hive (Carvell et al., 2006). The type and quantity of pollen and nectar gathered is indicative of the quality of the surrounding habitat and the importance of the codependent relationship between pollinators and the floral ecosystem (Dicks et al., 2015). As with any type of foraging behaviour, some food sources are preferable to others, be it as a result of ease of availability (Sipes and Tepedino, 2005) (complementarity of flower and pollinator morphology, for example), time and distance to forage, or the relative nutritional values (Liolios et al., 2015).

In order to characterise the plant species visited by pollinating bees, plant taxa contributing to the forage of a hive are generally identified by analysis of pollen in the resulting honey, or through physically tracking foraging bees over a set time period

(Carvell et al., 2006; Valentini et al., 2010). Molecular forage identification methods were popularised as DNA barcoding became more prevalent in plants (Newmaster et al., 2006), and were first driven, in part, by the validation of honey quality (Sivakesava and Irudayaraj, 2006) and identification of the region from which the honey was sourced (Burns et al., 2018). For example, honey sourced from specific plant forage, such as manuka honey from the tea tree plant (*L. scoparium*), can be sold at a much higher price than a non-specialist honey and as such validation methods are important to help reduce food fraud (Prosser and Hebert, 2017).

Until relatively recently, morphological identification of a subsample of pollen grains by microscopy was the preferred method of plant identification from honey (Von Der Ohe et al., 2004). The development of metabarcoding approaches, known colloquially as DNA barcoding (species identification through the analysis of DNA), (Hebert et al., 2003; Statnikov et al., 2013; Deiner et al., 2016), has benefits over methods based on morphology alone (Kohler, 2007). DNA barcoding methods have been used to identify organisms in diverse taxonomic groups including the bacteria, fungi, animals and plants (Janda and Abbott, 2007; Schoch et al., 2012; Hollingsworth et al., 2009; Hebert et al., 2003) and have been demonstrated to detect much greater diversity than morphological analysis of the same samples (Cowart et al., 2015) whilst providing comparatively accurate descriptions of a community (Lejzerowicz et al., 2015; Zimmerman et al., 2014). Where complex, mixed communities of microscopic, degraded, or otherwise cryptic taxa are considered (such as pollen in honey), metabarcoding using high-throughput, short-read sequencing technologies and associated computational analyses often provides the only tractable technique of describing the community (Deiner et al., 2017). Incremental improvements in affordability, computational methods and speed of data generation have allowed metabarcoding by next-generation sequencing (NGS) to become a mainstay of a wide variety of important analyses (Galan et al., 2017; Deurenberg et al., 2017; Borrell et al., 2017; Arulandhu et al., 2017).

The premise of DNA barcoding is of the rapid identification of a species or taxa by molecular comparison of a barcoding gene to a known reference sequence (ideally linked with a voucher specimen), based upon a quantifiable metric: similarity to the reference sequence (Pompanon et al., 2011; Coissac et al., 2012). The logical progression is from the barcoding of a single unknown template to DNA metabarcoding where the DNA of a mixed community, sometimes referred to as environmental DNA, is analysed in parallel ultimately generating a list of the taxa which have been detected in the sample (Moritz

and Cicero, 2004). One of the main advantages of this metabarcoding method is that after isolation of the target material (which would also apply to morphological methods of identification), and sequencing, the analysis is largely automated and computationally driven, rather than requiring expert, manual analysis of samples. This allows for the high-throughput analysis of samples, vastly increasing the data available for researchers, the speed of generation of results, and also for the standardisation of sample analysis improving repeatability between laboratories (Collins and Cruickshank, 2013). By making these sorts of community level analyses tractable to the conservation community, it is now possible to monitor, for example, the effect of pollution on marine sediment communities, survey the amphibians or fish present in freshwater ecosystems, or perform the parallel identification of the hundreds or thousands of individuals caught in an insect trap (Chariton et al., 2015; Ji et al., 2013; Valentini et al., 2015; Coissac et al., 2012).

There are many variables associated with metabarcoding NGS analysis that are known to bias the data produced. Sequencing technology, barcoding gene, *taq* DNA polymerase, number of PCR cycles, single or paired-end sequencing, and concentration of DNA template are all demonstrated to affect the sequencing error rate, rate of chimera production, taxa identified and relative proportions of taxa in sequence data (D'Amore et al., 2016; Dopheide et al., 2018; Nichols et al., 2018; Ramirez et al., 2018). Interspecies introgressions through backcrossing of hybrids can distort the taxa identifications generated by metabarcoding, and recent speciation events (on an evolutionary time frame), may generate two species which have not yet diverged sufficiently at a genetic level to achieve variability at the barcoding gene. Combined with copy number variation between the various genomes, taxonomic resolution of a marker varying between different taxa and interspecies imbalances in the rate of PCR amplification of a marker (based on size variation), these additional factors can cause inaccuracies in both taxa identifications, and also any inferences regarding the relative abundances of taxa within a sample (Ma and Li, 2015; Veltri et al., 1990; Coissac et al., 2012; Deiner et al., 2016).

Furthermore, there is not a standardised methodology for analysis of metabarcoding data; a particular problem for non-bacterial communities. Whilst different analysis packages should give similar outcomes, there are several very highly cited metabarcoding analysis packages, each built upon different fundamental methodologies (operational taxonomic units vs. exact sequence variants, for example) and the choice of analysis method likely introduces bias as significantly as any of the other variables described previously (Caporaso et al., 2010; Schloss et al., 2009; Callahan et al., 2016; Keegan

et al., 2016). As such, each of these biasing factors must proactively be considered and controlled, where possible, during the experimental design. Where control is impractical (controlling for sequencing technology would be prohibitively expensive in most cases), better understanding of the influence of the factors described here is invaluable. Whilst metabarcoding may hold great potential as a universal method of characterising the biota in a population from which a sample is drawn (Arulandhu et al., 2017), there is still a great deal to learn regarding taxa specific optimisation of variables in library preparation to produce the most reliable data.

A universal DNA barcode for plants has so far proved to be elusive, with a consensus emerging that multiple barcodes are required to obtain high confidence in taxa assignment (Newmaster et al., 2006; Bell et al., 2016). The Barcode of Life Data System (Ratnasingham and Hebert, 2007), an initiative designed to standardise barcoding approaches, endorses two DNA barcodes for use with plants: the chloroplastic ribulose bisphosphate carboxylase large chain gene (*rbcL*) and the plastidial maturase K (*matK*), (Bell et al., 2017; Yu et al., 2011). The *rbcL* barcode was chosen for the ease with which it can generally be amplified by PCR and *matK* for the resolution it affords due to rapid sequence changes in relatively recent evolutionary history (Hollingsworth et al., 2011). At >750bp, the amplified region of *matK* used for DNA barcoding is generally considered too large to be sequenced in its entirety, with suitable overlap for high-confidence assembly, by the current generation of next-generation high-throughput sequencers and is thus is more suited to Sanger sequencing approaches at present (Selvaraj et al., 2008). Advances in sequencing technology which are not read-length limited (for example devices from Oxford Nanopore Technologies) are now allowing longer DNA barcodes to be used, and non-targeted methods of metagenomics are also becoming common (Peel et al., 2019). For use on Illumina type sequencers, the nuclear Internal Transcribed Spacer 2 (*ITS2*) marker is rapidly becoming recognised as a suitable alternative for plant metabarcoding due to its discriminatory power and suitable length for high-throughput sequencing (Fahner et al., 2016; Laha et al., 2017). DNA barcodes from different genomes within the same organism (for example, the nuclear *ITS2* vs. the chloroplastic *rbcL*, both used for plant metabarcoding), are present in different quantities depending on the copy number variants (CNVs) present at each barcoding locus, which in turn vary by species, tissue type and physiological state (Ma and Li, 2015; Veltri et al., 1990). The proportions of chloroplastic and nuclear DNA vary within a sample meaning that each barcoding gene is effectively describing a different

community within the DNA extraction (D'Amore et al., 2016), and therefore inconsistencies in relative abundance of taxa between these competing descriptors are likely. Given the inherent biases in NGS metabarcoding, here we refer to the debate (Mallott et al., 2018; Hollingsworth et al., 2011; Sickel et al., 2015) on the most suitable plant barcoding gene, and provide some empirical evidence of the performance of three markers. We perform direct comparisons between two barcoding genes, consisting of an *rbcL* marker and two variants of *ITS2*, taken from the literature. These three markers, represent two plant specific barcoding markers and one general barcoding marker, and allow us to investigate the impacts of marker choice on the confidence and validity of honey metabarcoding results.

The aim of our study is to answer the following core questions:

1) How confidently are we able to identify plants from a pollen sample?

2) How similar are the results from the three markers and two marker families?

3) Does a pair of markers allow for greater confidence in the description of a community?

4) Which pair of markers provides the highest confidence in community description?

5.4 Materials and Methods

5.4.1 Sampling and Molecular Biology

Honey samples were collected from 15 *Apis mellifera* hives in Greater Manchester and the surrounding area; the majority being urban hives (Figure 5.1 and Table 5.1). Following a published methodology (Hawkins et al., 2015), DNA was extracted from 40g honey (four parallel extractions of 10g each, where possible) using a modified protocol for the DNeasy Plant Mini Extraction Kit (Qiagen, Hilden, Germany). Briefly, four sets of 10g honey were each diluted in 25ml molecular biology grade H_2O and incubated at 65°C with periodic shaking. Once completely dissolved, the honey was centrifuged at 15,000 x g for 30 minutes with the supernatant discarded and the pellet suspended in 400µL AP1 buffer (DNeasy Plant Mini Extraction Kit) with proteinase K (Bioline, London, UK) added at a concentration of 20mg/ml. Samples were processed with a Retsch MM400 mixer mill (Retsch, Haan, Germany), (3mm tungsten carbide beads, four one minute cycles at 30 Hz) before a further incubation at 65°C. Each set of four parallel extractions were pooled in a single DNeasy Plant Mini kit spin column and the extraction continued as directed by the manufacturer's manual with the exception of omissions of the QIA shredder column

step and the second wash stage. Extracted DNA was frozen for long-term storage at -80°C. For amplification, extracted DNA was diluted 1:5 in molecular biology grade H_2O. Three barcoding regions were amplified using PCR: the chloroplastic *rbcL* gene (Hawkins et al., 2015; de Vere et al., 2012) and two variants of the nuclear, internal transcribed spacer 2 (*ITS2*) gene (Cheng et al., 2016), (Table 5.2). One *ITS2* marker was selected for plant specificity and one for universal *ITS2* amplification. Herein they are referred to as *ITS2* plant (*ITS2p*) or *ITS2* universal (*ITS2u*), respectively. Amplifications were carried out using a modified version of the standard two-step Illumina 16S library preparation protocol. The first round of PCR amplifies the locus of interest and adds Illumina specific sequencing adapters to the $5'$ end of each strand of the amplicons. Reactions consisted of 2µL template DNA, 12.5µL 2x KAPA HotStart Ready Mix (Roche, Basel, Switzerland), 0.5µL of each primer (10µM) and 9.5µL molecular biology grade H_2O. Purified amplicons had Illumina sequencing adaptors (Illumina, California, USA) added in a second round of PCR. This further amplification and the next-generation sequencing were performed at the Centre for Genomic Research at The University of Liverpool. Sequencing was performed using an Illumina HiSeq 2500 platform (Illumina, California, USA) with the 2x 250bp rapid, V2 chemistry.

5.4.2 Sequence Data Analysis and Quality Control

Raw sequence data was demultiplexed by Illumina index sequence into individual sample/barcoding gene combinations (three data sets per biological sample). Primer sequences were removed from sequencing reads using the program Cutadapt (Marcel, 2011), accounting for degeneracies within primer sequences. Up to two instances of each primer were removed from each read and pairs of reads which did not contain both primer sequences were discarded. Pairs of reads were assembled using the QIIME2 (Caporaso et al., 2010) tool "join_paired_ends.py", with the default settings. Operational taxonomic units (OTUs) were picked using the QIIME clustering tool "pick_open_reference_otus.py" with either an *ITS2* (Sickel et al., 2015) or *rbcL* (Bell et al., 2017) curated sequence database provided as appropriate, and clustering performed around 97% sequence similarity, for all markers. Analyses of species discrimination at various clustering similarity thresholds, show that in UK flora, 97% clustering of *ITS2* sequence data gives high to moderate species level discrimination in the majority of orders of plants (Moorhouse-Gann et al., 2018) compared to more stringent thresholds. Clustering of *rbcL* data does not require such stringent similarity,

with 91%-92% sequence similarity discovered to give the best species level discrimination in diatoms, and 97% sequence similarity a commonly used threshold (Tapolczai et al., 2018; Erickson et al., 2017). To minimise further confounding factors in our comparison of the two markers, we elected to use the same similarity threshold for both markers, and chose a value representing an acceptable middle-ground for both datasets. Binary "biom" files were converted to text format using biom-format (McDonald et al., 2012).

Taxonomy assignment was performed via the default QIIME methodology that implements the UCLUST algorithm (Edgar, 2010) to find the closest match for an OTU in the reference database. For comparison, unassigned OTUs were further processed using the BLAST algorithm against the curated database (100% sequence match along 75% of the OTU sequence required for assignment), but these secondary assignments were not subsequently used for the current community analysis (Table 5.3). The removal of chimeric sequences from the raw data was initially found to be prohibitively computationally intensive. Therefore the OTU clustering process was first performed on raw data containing potential chimeric sequences and the relevant reference database filtered to just those entries which were found in the resulting data set after taxonomy had been assigned. Chimera detection using UCHIME2 (Edgar, 2016) was then performed on the raw data against these reduced databases, allowing chimera removal to run in an acceptable time frame. The *ITS2* database was reduced to 629 species and the *rbcL* database reduced to 771 species. As expected, the *ITS2u* marker amplified both plant taxa and non-plant taxa with the majority of data classified in the fungal kingdom. Any OTUs which were not assigned a taxonomy within the plant kingdom, in any marker, were removed from further analysis. Relative abundance counts were calculated and very low frequency OTUs (<0.001%) removed.

Plant species assignments were checked for plausibility of their presence in the UK against several databases: the Royal Horticultural Society Horticultural Database (Royal Horticulutural Society, 2014) based upon the BG-BASE database (version 7.3, accessed 2018), the Biological Records Centre Atlas of the British and Irish Flora (2018 version, accessed 2018) (Biological Records Centre, 2019), and the Plants For A Future database (Plants For A Future, 2019), (2018 version, accessed 2018). Species missing from plant databases were checked for availability in online, UK based garden centres and assigned UK plausibility accordingly. As the sampling sites were mainly in urban and suburban areas, forage was expected to be sourced primarily from gardens. Therefore, databases of natural flora in the UK were unlikely to be sufficient to

categorise the full range of plant diversity encountered by the bees (de Vere et al., 2012). This extensive, but intrinsically non-exhaustive, list of plausible UK plant species was used in the analysis of all three metabarcoding markers.

Given well documented difficulties in achieving species level resolution with plant metabarcoding markers (Fahner et al., 2016) we implemented a method to increase our confidence in species level assignments. Assuming the logical position that a species appearing in the data sets of two independent markers (*rbcL* and an *ITS2* marker) constitutes a higher quality assignment, we implemented a system of taxa cross-validation for species level assignment. Two markers, *rbcL* and *ITS2u* were chosen due to this pair of independent markers targeting the most diverse group of plant taxa, and thus being likely to capture the most diversity in the sample. Genus level assignments were retained in all three markers, with no cross-validation performed. Therefore, confidence of taxonomic assignment was categorised and determined by the following method:

1) If a species was identified by both the *rbcL* marker and the *ITS2u* marker, it was assigned species level identification.

2) If a species was identified by either *rbcL* or *ITS2u* marker and it was the only species in that genus in the entire data set, it was assigned species level identification.

3) Otherwise, it was assigned genus level identification.

Those OTUs that were plausible in the UK and achieved confident species level assignment were retained in the species level data set. Plausible genera which did not contain any plausible species were removed. All OTUs were collapsed into genus, family and order and the total for each taxonomic classification calculated. Provided a genus contained $\geq$1 species plausible as being present in the UK, every OTU in that genus was retained in the genus level data set and data sets at higher taxonomic levels.

5.4.3 Statistical Analyses

Relative abundance counts were transformed using the Wisconsin double standardisation method in R 3.4.4 (R Development Core Team, 2008). The R package "vegan" (Oksanen et al., 2018) was used to calculate Bray-Curtis (BC), (Bray and Curtis, 1957) dissimilarity indices between community matrices, perform ANOSIM analysis and to calculate measures of alpha diversity. Dissimilarity indices were also calculated between the three distinct descriptors of each plant community (biological sample) at order, family and genus, and between *rbcL* and *ITS2u* at species levels.

Pairwise Mantel (Mantel, 1967) tests were performed upon the genetic distance matrices of the three marker data sets, encompassing all 15 samples. Calculations were again made at four taxonomic levels for comparison. Bonferroni correction for multiple tests was performed using the p.adjust function in R.

5.5 Results

5.5.1 Quality Control and Taxon Assignment

The rate at which raw reads were removed by quality control filters was similar between the three markers, but deviated at the point where OTUs were generated. At a 97% similarity threshold, the number of OTUs generated in the *ITS2* markers (*ITS2p*: 1141, *ITS2u*: 1133) was much greater than in *rbcL* (231), (Figure 5.2). When 97% OTUs which had been assigned identical taxa were collapsed to unique taxa, *rbcL* retained 87.4% of the original OTU count whilst *ITS2u* retained just 60.5%, indicating that analysis of *ITS2* data may benefit from a lower similarity threshold during clustering. Whilst *rbcL* produced many fewer 97% OTUs, it retained a greater proportion after filtering for UK plausibility. On average *rbcL* retained 88.6% of unique OTUs after filtering to plausible genera, compared to 75.4% in *ITS2u* and 77.0% in *ITS2p*.

Our assignment algorithm allowed us to assign 42.9% of total *rbcL*, and 45.3% total *ITS2u* reads to the species level, and 43.9% *rbcL* reads and 42.9% *ITS2u* reads to genus level only. Of the two markers studied here, *rbcL* appears to have the lowest rate of false positive identification, based on our methods of taxa confirmation. In *rbcL*, 11.4% OTUs, and in *ITS2u*, 24.6% of OTUs were removed as they were assigned to implausible taxa. These OTUs accounted for 13.1% of *rbcL*, and 12.4% of *ITS2u* raw sequence reads. Of all false-positive OTUs filtered from the data set, 91.1% were at a relative abundancy of <1% and 75.1% were at a relative abundance of <0.1%, indicating that low frequency OTUs were very likely to be generate false positive assignments.

5.5.2 Sample Diversity

Shannon's diversity index was higher in *rbcL* than either *ITS2* markers, which were very similar (mean Shannon's diversity index: *rbcL*: 1.22 [95% CI: 0.93;1.50], *ITS2p*: 0.75 [95% CI: 0.35;1.14], *ITS2u*: 0.76 [95% CI: 0.45;1.07]), with the two *ITS2* descriptions highly correlated (Spearman's rank correlation R=0.83), and *rbcL* and *ITS2* descriptions

less so (*rbcL* vs *ITS2u* R=0.45, *rbcL* vs *ITS2p* R=0.69), (Figure 5.3.) Species evenness results also gave the same descriptions of alpha-diversity, and correlation between markers: *rbcL* species evenness: 0.56 [95% CI: 0.45;0.67], *ITS2p* species evenness: 0.20 [95% CI: 0.11;0.29], *ITS2u* species evenness: 0.21 [95% CI: 0.14;0.28]), with the pair of *ITS2* again the most highly correlated: Spearman's rank correlation: *ITS2u* vs. *ITS2p* R=0.75, *rbcL* vs. *ITS2u* R=-0.18, *rbcL* vs. *ITS2p* R=-0.25, (Figure 5.4). Analysis of relative counts of genera detected in each family identified, shows distinct variation between the markers (Figure 5.5) indicating fundamental differences between the plant communities described by each marker. The *ITS2u* marker is seen to identify many more unique genera than both *rbcL* and *ITS2p*, across a broader range of families. However the most apparent difference in family representation is between *rbcL* and *ITS2u*. This analysis shows some familial bias inherent in marker choice with many families relatively highly represented in one or more markers, but much lower, or absent, in others. The average number of genera per family in each marker was 2.24 in (*rbcL*), 4.65 (*ITS2p*), and 5.30 in (*ITS2u*).

5.5.3 Statistical Comparison of Communities

Both the Bray-Curtis (BC), (Bray and Curtis, 1957) and Jaccard (Jaccard, 1908) dissimilarity indices were used to compare pairs of community matrices describing each plant community, at order, family and genus taxonomic levels. Only *rbcL* and *ITS2u* data was compared at the species level, due to our taxon assignment method (Figure 5.6 and Figure 5.7). Ranging from zero (containing the same taxa) to one (containing different taxa), the results show that these barcoding genes describe increasingly dissimilar communities as greater taxonomic resolution is achieved, with the most notable increase in dissimilarity occurring when genus level identifications are made. Mantel tests on both BC and Jaccard values (Table 5.4) among community matrices showed that the pair of *ITS2* markers produce the most similar community descriptions with other pairs of markers differing significantly, with *rbcL* and *ITS2u* being the least similar. The BC dissimilarity index considers relative abundance counts, whilst the Jaccard dissimilarity index counts presence-absence data only. We have established that the different barcoding genes used appear to have familial biases (Figure 5.5) in terms of taxa amplified (*rbcL* vs *ITS2*), and this variation is seen in calculations of Jaccard's index, where these pairs of markers describe quite different communities. We would expect the BC index between the pair of *ITS2* markers to be less impacted by variation in relative abundance of taxa as both markers target the same gene, and so copy number variation will have less of

an effect. We do see less of an increase in dissimilarity in the BC index, compared to Jaccard, however both measures still describe quite dissimilar communities, pointing to other factors impacting variation in taxa assignment and relative abundances, possibly linked to variation in PCR efficiency across taxonomy or other stochastic factors.

Analysis of similarities (ANOSIM) analysis also revealed a congruent result: significant differences in community composition between *rbcL* and *ITS2p* (ANOSIM R statistic = 0.3083, p = 0.001) and *rbcL* and *ITS2u* (R = 0.2085, p = 0.002), but marginally significant between the pair of *ITS2* markers (R = 0.0538, p = 0.055). Core species or genera in each biological sample (defined as species/genus which occurred in the *rbcL* data set and *ITS2u* data set), in many cases was limited to just a single species and ranged from one to five genera.

We recorded notable variation in the proportion of reads per marker which were able to be assigned a taxonomy using each of two methods tested. For both *rbcL* and *ITS2u*, most reads were assigned using the UCLUST algorithm (74.8% and 76.5%, respectively). Conversely, most *ITS2p* reads were able to be assigned a taxonomy using the BLASTn method (Table 5.3). There is very clear, marker specific variation in the effectiveness of different algorithms to match OTUs to their respective reference databases.

5.6 Discussion

We have used next-generation DNA sequencing and a bioinformatics workflow to describe plant taxa in DNA extracted from pollen in honey. We saw extremely high rates of false positive results, highlighting the ease with which over-interpretation of plant metabarcoding data can occur. We recommend the implementation of a method by which the plausibility of results are tested, particularly if attempting to establish species level identifications. A database of plausible taxa in the ecosystem is extremely valuable for quality control and should be generated, and/or evaluated with the highest possible level of stringency (Schloss and Westcott, 2011; Holovachov et al., 2017). Through the parallel analysis of data sets generated by three plant metabarcoding markers, we were able to compare their community descriptions, and devise a method of validating taxa assignments. We found that competing descriptions of a plant community were highly divergent, with statistical tests showing significant differences in many instances. Utilising a pair of divergent DNA markers increases the range of taxa which can be amplified, maximising the diversity of plants able to be described in the sample (Fahner

et al., 2016). Furthermore, taxa which appear in two divergent markers can be assigned a higher level of confidence, and a higher taxonomic resolution. For these purposes, we recommend parallel analysis of plant community samples with *rbcL* and *ITS2u* markers. This pair of markers were chosen as they are amplified regions of two independent genes and therefore the diversity of taxa amplified, and the confirmatory power of taxa identified by both markers, is maximised. Of the two *ITS2* variants tested here, the universal variant *ITS2u* had a lower rate of false positive identifications, indicating fewer incorrect assignments, however further investigations into the accuracy of the plausible assignments will be required. Presently, we are not able to say whether a plausible taxonomic result is a true-positive, however taxa identified by these two independent markers are logically the most likely. The amplicon size variation in *ITS2p*, is slightly greater than in the universal variant and may, in some taxa, produce an amplicon which cannot be sufficiently assembled after paired-end sequencing using 2 * 250bp chemistry. This would result in the technical exclusion of taxa with the longest section of *ITS2* which falls between the two primers. As the *ITS2u* PCR amplicon is shorter, and falls more comfortably within the size of product suitable for sequencing using 2 * 250bp sequencing, we would further argue for its utility as the more suitable barcoding marker.

The speed and relative ease by which mixed biological samples can be "barcoded" using next-generation molecular techniques has been revolutionary to many fields of biology (Cristescu, 2014). Metabarcoding provides a much needed high-throughput method of taxonomic assignment but is susceptible to mis- or over interpretation without due care (Elbrecht et al., 2017). It is well established that almost every stage of the preparation of a DNA library for NGS metabarcoding introduces bias into the results (D'Amore et al., 2016), however there is more to be done to better understand how these biases can be controlled to provide greater confidence in taxonomic assignments. Experiments allowing direct comparisons between library preparation treatments are therefore invaluable. In the present study we performed parallel analysis of pollen communities using three barcoding markers from the literature (Hawkins et al., 2015; Cheng et al., 2016) to better understand the variation in results attributable to barcoding gene. Our study tested just three potential plant metabarcoding markers which amplify regions of two genes. *rbcL* and *ITS2* are both popular markers for plant metabarcoding due to their properties making them suitable for analysis upon the current generation of short read sequencers (Hawkins et al., 2015; de Vere et al., 2012), however there are several other highly cited metabarcoding markers in the literature

including *matK* and *trnL*. Further comparisons between metabarcoding markers, across a range of types of flora will continue to be useful to and informative to the plant metabarcoding community (Fahner et al., 2016; Mallott et al., 2018).

Our results suggest that a single DNA barcoding marker does not provide a reliable method of identifying plant species in a mixed sample. This conclusion is driven by the high variability seen between community data sets generated by pairs of markers. We have concluded that using two divergent markers for analysis allows for higher confidence calls to be made where a taxon is identified by both markers, although resolution of taxonomic assignment remains an issue (Fahner et al., 2016). However, not all combinations of cross-verification are equal. Complementarity of pairs of DNA markers has been discussed widely in respect to analysis of plant communities, with breadth of amplification an important factor. Length of DNA barcode, limited by degradation of target DNA and the read-length limitations of DNA sequencers, limits the resolution afforded by a single marker. Additional power, through the use of complementary markers has been shown to be an effective method. For example, the addition of a second short marker (a region of *ITS* combined with the *trnL P6 loop*) allowed for greater resolution in the Sapotaceae family, beyond that which was available using plastid markers only (Yoccoz et al., 2012). A promising strategy for the combination of markers is to use a marker with high accuracy and taxonomic performance with a second marker chosen for the wide breadth of amplification. Using such a strategy allows for an increase in the likelihood of detecting, and identifying the broadest range of taxa in a sample. For analysis of plant communities *rbcL* and *ITS2* are strong candidates for this approach and have the added benefit of being supported by the development of high quality reference databases (Fahner et al., 2016; Laha et al., 2017). Taxonomic assignments generated by each of the *ITS2* markers are more closely correlated with one another due to them being variants of the same marker. They each amplify the same region of the *ITS2* gene and differ only in their use of degenerate bases in the primer design to widen the range of amplifiable taxa, in the case of the universal variant. As such cross-verification between these two markers is not as robust as cross-verification between *rbcL* and either *ITS2* marker. To avoid introducing a bias towards *rbcL* classifications, only *rbcL* assigned taxa which were confirmed in the universal variant of *ITS2* (*ITS2u*) were given confident species level status. During community analysis, this pair of markers were found to be the most divergent, giving the widest scope for possible taxa identification. This divergence also means we can imply higher confidence in their complementary taxa

assignments, which also informed their selection as the optimum pair. The logic of this cross-validation method is in line with other studies which recommend that for metabarcoding of plants, multiple barcoding genes should be used concurrently (Newmaster et al., 2006; Ratnasingham and Hebert, 2007; Fahner et al., 2016).

The different barcoding markers investigated did not produce similar descriptions of the same biological sample, with increasing dissimilarity at greater taxonomic resolution. The universal nature of the *ITS2u* marker resulted in the amplification and sequencing of the fungal component of the honey, data from which was removed during the bioinformatics analysis. In this instance, a large capacity sequencer was used (Illumina HiSeq) and this generation of data not useful to the present study, did not represent a significant set back as the quantity of plant data generated alongside the fungal, was of a sufficiently large volume to detect even relatively low abundance plant taxa. However, if either a smaller capacity sequencing platform were used, species rich samples analysed (particularly those with a large fungal component) or research was directed particularly towards low frequency plant taxa, then the co-generation of fungal sequence data may present an issue in the analysis of those samples. After the removal of non-plant taxa from the *ITS2u* data set, *rbcL* was consistently very different from the *ITS2u* data, as exhibited by very high BC dissimilarity indices and Mantel tests upon community matrices. Cross-validation allowed higher confidence in some taxonomic assignments, allowing some species level identifications to be made. However, given the high levels of disparity between results generated by pairs of markers, this results in very few high confidence species assignments per sample. In some samples we recorded that very high proportions of sequence reads (sometimes >99%) could be attributed to a single species (often *Impatiens glandulifera*). However, low plant diversity in bee forage is common (Donkersley et al., 2017) and low diversity in data is not therefore necessarily indicative that taxa are being missed. There are several possible explanations as to why these two barcoding markers would give significantly different community descriptions of the same biological sample. As *rbcL* is a chloroplastic gene and *ITS2* a nuclear gene, it is likely that the relative proportions of total chloroplastic DNA and total nuclear DNA differ significantly in the DNA extractions (Rauwolf et al., 2010). The ratio between chloroplastic genomic DNA and nuclear genomic DNA also likely varies for each plant species in the community. This results in variation in starting template concentration (D'Amore et al., 2016), and also an imbalance in the "true" pollen communities represented by each type of genome. Furthermore, inheritance of the chloroplast is

bi-parental in some species and occurs through a single parent in others; however it is most commonly inherited through the maternal lineage (Ellis et al., 2008; Bell et al., 2016). We found a high rate of false positive errors overwhelmingly associated with large numbers of very low frequency OTUs being attributed to non UK plausible taxa. From our own analysis, we argue that a minimum OTU threshold of 1% relative abundance would be advisable as the false positive taxa were generally those present in abundances below this threshold. These low frequency false positive OTUs may be associated with nucleotide errors generated during sequencing or statistical artefacts associated with clustering reads in highly divergent marker genes (Ma et al., 2019; Hathaway et al., 2018; Callahan et al., 2016). Alternatively, software packages are available which facilitate a different method of analysing metabarcoding data, which does not require clustering around a similarity threshold. One of these, DADA2 (Callahan et al., 2016), includes a method of correcting sequencing errors (the Divisive Amplicons Denoising Algorithm) in the data, and may be an alternative, or preferable, method of analysing data of this sort. Alternatively the high false positive rate may be due to discriminatory resolution being limited by the length of sequencing read currently available (Maloukh et al., 2017; Moorhouse-Gann et al., 2018). We used the 2 * 250bp sequencing chemistry to sequence amplicons, giving a theoretical maximum read length of approximately 450 bp, once paired-end read assembly has occurred (which requires a significant amount of sequence overlap for high quality assemblages to be generated). This limitation of read length to a relatively short amplicon (compared to Sanger sequencing which may give at least 1,500bp in high quality, assembled data, or third generation sequencers which do not have a maximum read-length) will very likely have restricted the taxonomic resolution of each marker. A trade off must be found where one is able to use high-throughput sequencing methods (I.e. as opposed to Sanger sequencing), but still find molecular markers with sufficient power to resolve taxa at an appropriate level.

The majority of OTUs removed were found to be in plausible genera, but were assigned to implausible species. Thus indicating that these markers may provide sufficient accuracy to perform genus level assignment, but caution should be exercised where species level results are to be interpreted. Attempting genus level assignment only would therefore reduce the false positive rate significantly. Where species within a genus are separated by a very small percentage of nucleotide substitutions, the read-length and divergence of markers may make species resolution impossible using these markers and current generation of sequencers (Fahner et al., 2016). Whilst we recommend the

removal of very low frequency OTUs, as they are the most likely to be false-positives, this does however raise the chance of removing very low frequency taxa actually present in the sample. Using the markers in this study, metabarcoding of plant DNA in honey is not currently sufficiently developed to confidently detect very rare taxa.

Sequence redundancy is inherent in a large NGS metagenomics data set and is a strength often harnessed via sequence clustering and the production of OTUs (Caporaso et al., 2010). The error rate of modern sequencing platforms varies by technology (Quail et al., 2012), but clustering allows the removal of sequencing artefacts and errors through the generation of consensus sequences (Li et al., 2012). However, there is the potential to miss fine scale variation in sequence which is masked by the similarity threshold used for clustering (Callahan et al., 2016). In this comparative study, the generation of OTUs around a 97% sequence similarity threshold revealed a greater level of sequence variation in the *ITS2* region compared to *rbcL*, as evidenced by the greater number of OTUs produced using *ITS2* markers. This reflects the fact that the amplified section of the *rbcL* gene shows much less sequence variation than *ITS2*, as expected due to *rbcL* being under high selective pressure due to its role in photosynthesis, and *ITS2* being a non-coding region (Kress and Erickson, 2007). The optimal similarity threshold around which sequences should be clustered appears to be barcoding gene dependent (Schloss and Westcott, 2011; Holovachov et al., 2017). Duplication of taxonomically identical 97% OTUs was very high in all both markers but was much higher in *ITS2u*. After removal of false positive assignments, *rbcL* retained a higher proportion of OTUs than *ITS2u*, and therefore in this instance has the lowest effective false positive rate when considering 97% OTUs. Relatively recently, metabarcoding bioinformatics methods which do not rely upon the clustering of similar sequences into OTUs, but rather employ analysis of exact sequence variants (ESV) have been developed (Callahan et al., 2016). Such methods potentially have greater taxonomic resolution due to the ability to detect low frequency nucleotide substitutions in sequence data, as every sequence is analysed individually (Callahan et al., 2017). Drawbacks to the ESV methodologies include a greater sensitivity to sequence error and sequencer generated error, and the production of excessive variation in otherwise low diversity samples. Our recommendations to remove any OTUs with less than 1% relative abundance demonstrate that in this instance, detecting low frequency variants in the data was not a priority. The benefits of the generation of consensus sequences and data reduction are a valid approach, with the optimum data analysis method likely depending upon sample diversity and the aims of the study.

We have established that the validation of taxa assigned by the chosen analysis method against a high quality, appropriate database is critical. In the British Isles we have an extremely well characterised and barcoded native flora (Ratnasingham and Hebert, 2007), but in reality the range of plants visited by bees and other pollinators is often much more diverse. Gardens stocked with exotic plants very often provide valuable resources to pollinators at times of the year when native plants are less productive (Donkersley et al., 2017). For studies which are particularly focussed on urban and suburban ecosystems, it is important to adjust the criteria by which false positives are identified. Where bees access cultivated gardens, a simple filter which only passes native plant species will not suffice. A comprehensive database of both native, and non-native cultivars plausibly detected in the ecosystem is essential.

Our taxonomy assignment method was designed to allow taxonomic assignment of as many OTUs as possible, by initially utilising the UCLUST method and secondarily applying BLAST. We ultimately decided against using these secondary assignments as they would have introduced an extra variable (assignment method). There is a very clear discrepancy between the two *ITS2* markers as to which assignment method was able to assign the most OTUs This highlights the need for an informed decision to be made as to the choice of classifier, and is in line with other studies which have investigated classifiers in more details (Holovachov et al., 2017). It is therefore important that multiple methods are tested and compared, and that in the case where multiple markers are used in an experiment, a uniform analysis approach is not necessarily applied to every marker without consideration of this variation.

The different markers we employed here, and other metabarcoding markers, vary in their ability to report the correct taxonomy of a sample. The resolution of a marker is influenced by the inter-taxa variation, within the amplified region of the gene and the ability of the sequencer and bioinformatics processes to detect that variation. Whilst species level identifications are clearly the goal, the resolution of metabarcoding markers and the current generation of next-generation sequencers do not easily give species level identifications in plants. Attempting family or genus level identifications only, may give greater confidence in the taxonomic results, by trading off against specificity. Markers which amplify genes from different genomes (nuclear, chloroplastic, plastidial) are effectively describing different communities, despite the DNA being extracted from the same sample. Variation in the concentration of starting template, or copy number variation, differs between each of these genomes and is altered by various physiological

factors, meaning that variation in community descriptions of taxa, and of their relative abundance, are to be expected as a technical factor associated with plant metabarcoding (Bell et al., 2017; Yu et al., 2011; Hollingsworth et al., 2011; Ma and Li, 2015; Veltri et al., 1990).

The results generated here clearly demonstrate difficulties which face researchers interested in metabarcoding of pollen, and analysis of communities more generally, and will be useful for anyone considering the choice of which barcoding genes to use in their research. We were able to confidently assign taxa to plants detected in pollen samples, giving higher confidence assignments based upon our cross-validation method. The results from the three markers, and two marker families, which we employed for this analysis were dissimilar, however we were able to use this divergence as a strength to imply higher confidence in taxa assignment generated by a pair of highly divergent markers. By choosing the most divergent pair of markers for the cross-validation method, we were able to access both the broadest range of amplified taxa, and assign them the highest confidence.

5.7 Acknowledgements

We are very grateful to all the beekeepers who kindly collected, and donated honey samples from their hives, as well as the Manchester and District Beekeepers Association (MDBKA). Special thanks also go to Yan Guo who assisted with much of the laboratory work and processing of samples. Funding for the sample collection, preparation and sequencing was received from a Daphne Jackson Trust Research Fellowship (Daphne Jackson Trust) grant in Latha R. Vellaniparambil's name. Funding for Graeme Fox's PhD project comes from Manchester Metropolitan University.

5.8 Conflicts of Interest

The authors declare that they have no conflict of interest.

References

Arulandhu, A., Staats, M., Hagelaar, R., Voorhuijzen, M., Prins, T., Scholtens, I., Costessi, A., Duijsings, D., Rechenmann, F., Gaspar, F., Crespo, M., Holst-Jensen, A., Birck, M., Burns, M., Haynes, E., Hochegger, R., Klingl, A., Lundberg, L., Natale, C., Niekamp, H., Perri, E., Barbante, A., Rosec, J., Seyfarth, R., Sovová, T., Moorleghem, V., van Ruth, S., Peelen, T., and Kok, E. (2017). Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. *Gigascience*, 6(10):1–18.

Bell, K., de Vere, N., Keller, A., Richardson, R., Gous, A., Burgess, K., and Brosi, B. (2016). Pollen DNA barcoding: current applications and future prospects. *Genome*, (59):629–640.

Bell, K., Loeffler, V., and Brosi, B. (2017). An *rbcL* reference library to aid in the identification of plant species mixtures by DNA metabarcoding. *Applications in Plant Sciences*, 5(3):1600110.

Biological Records Centre (2019). Biological Records Centre Atlas of the British and Irish Flora. Accessed: 16/01/2019.

Bloom, E., Northfield, T., and Crowder, D. (2019). A novel application of the price equation reveals that landscape diversity promotes the response of bees to regionally rare plant species. *Ecology Letters*, 22(12).

Borrell, Y., Miralles, L., Huu, H., Mohammed-Geba, K., and Garcia-Vazquez, E. (2017). DNA in a bottle - rapid metabarcoding survey for early alerts of invasive species in ports. *PLOS ONE*, 12(9):e0183347.

Bray, J. and Curtis, J. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27:325–349.

Burns, D., Dillon, A., Warren, J., and Walker, M. (2018). A critical review of the factors available for the identification and determination of mānuka honey. *Food Analytical Methods*, 11(6):1561–1567.

Callahan, B., McMurdie, P., and Holmes, S. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *Multidisciplinary Journal of Microbial Ecology*, 11(12):2639–2643.

Callahan, B., McMurdie, P., Rosen, M., Han, A., Johnson, A., and Holmes, S. (2016). DADA2: High resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7):581–3.

Caporaso, J., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F., Costello, E., Fierer, N., Pena, A., Goodrich, J., Gordon, J., Huttley, G., Kelley, S., Knights, D., Koenig, J., Ley, R., Lozupone, C., McDonald, D., Muegge, B., Pirrung, M., Reeder, J., Sevinsky, J., Turnbaugh, P., Walters, W., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336.

Carvell, C., Roy, D., Smart, S., Pywell, R., Preston, C., and Goulson, D. (2006). Declines in forage availability for bumblebees at a national scale. *Biological Conservation*, 132(4):481 – 489.

Chariton, A., Stephenson, S., morgan, M., Steven, A., Colloff, M., Court, L., and hardy, C. (2015). Metabarcoding of benthic eukaryote communities predicts the ecological condition of estuaries. *Environmental Pollution*, 203:165–174.

Cheng, T., Xu, C., Lei, L., Li, C., Zhang, Y., and Zhou, S. (2016). Barcoding the kingdom plantae: PCR primers for *ITS* regions of plants with improved universality and specificity. *Molecular Ecology Resources*, 16(1):139–149.

Coissac, E., Riaz, Y., and Puillandre, N. (2012). Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, 21(8):1834–47.

Collins, R. and Cruickshank, R. (2013). The seven deadly sins of DNA barcoding. *Molecular Ecology Resources*, 13(6):969–75.

Cowart, D., Pinheiro, M., Mouchel, O., Maguer, M., Grall, J., Mine, J., and Arnaud-Haond, S. (2015). Metabarcoding is powerful yet still blind: A comparative analysis

of morphological and molecular surveys of seagrass communities. *PLOS ONE*, 10(2):e0117562.

Cristescu, M. (2014). From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology and Evolution*, 29(10):566–571.

D'Amore, R., U.Z., I., Schirmer, M., Kenny, J., Gregory, R., Darby, A., Shakya, M., Podar, M., Quince, C., and Hall, N. (2016). A comprehensive benchmarking study of protocols and sequencing platforms for *16S* rRNA community profiling. *BMC Genomics*, 17(55).

de Vere, N., Rich, T., Ford, C., Trinder, S., Long, C., Moore, C., Satterthwaite, D., Davies, H., Allainguillaume, J., Ronca, S., Tatarinova, T., Garbett, H., Walker, K., and Wilkinson, M. (2012). DNA barcoding the native flowering plants and conifers of Wales. *PLOS ONE*, 7(6):e37945.

Deiner, K., Bik, H., Machler, E., Seymour, M., Lacoursiere-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D., de Vere, N., Pfrender, M., and Bernatchez, L. (2016). Environmental DNA metabarcoding: Transforming how we survey animal and plant comuniities. *Molecular Ecology*, 26(21):5872–5895.

Deiner, K., Bik, H., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D., Vere, N., Pfrender, M., and Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, 26(21):5872–5895.

Department for Environment Food and Rural Affairs (2014). The national pollinator strategy: for bees and other pollinators in England. United Kingdom.

Deurenberg, R., Bathoorn, E., Chlebowicz, M., Couto, N., Ferdous, M., Garcia-Cobos, S., Kooistra-Smid, A., Raangs, E., Rosema, S., Veloo, A., Zhou, K., Friedrich, A., and Rossen, J. (2017). Application of next-generation sequencing in clinical microbiology and infection prevention. *Journal of Biotechnology*, 243:16–24.

Dicks, L., Baude, M., Roberts, S., Phillips, J., Green, M., and Carvell, C. (2015). How much flower-rich habitat is enough for wild pollinators? answering a key policy question with incomplete knowledge. *Ecological Entomology*, 40(S1).

Donkersley, P., Rhodes, G., Pickup, R., Jones, K., Power, E., Wright, G., and Wilson, K. (2017). Nutritional composition of honey bee food stores vary with floral composition. *Oecologia*, 185(4):749–761.

Dopheide, A., Xie, D., Buckley, T., Drummond, A., and Newcomb, R. (2018). Impacts of DNA extraction and PCR on DNA metabarcoding estimates of soil biodiversity. *Methods in Ecology and Evolution*, 10(1).

Edgar, R. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461.

Edgar, R. (2016). UCHIME2: improved chimera prediction for amplicon sequencing. *bioRxiv*, 074252.

Elbrecht, V., Vamos, E., Meissner, K., Aroviita, J., and Leese, F. (2017). Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution*, 8(10):1265–1275.

Ellis, J., Bentley, K., and McCauley, D. (2008). Detection of rare paternal chloroplast inheritance in controlled crosses of the endangered sunflower *Helianthus verticillatus*. *Heredity*, 100(6):574–580.

Erickson, D., Reed, E., Ramachandran, P., Bourg, N., McShea, W., and Ottesen, A. (2017). Reconstructing a herbivore's diet using a novel *rbcL* DNA mini-barcode for plants. *AoB Plants*, 9(3).

Fahner, N., Shokralla, S., Baird, D., and Hajibabaei, M. (2016). Large-scale monitoring of plants through environmental DNA metabarcoding of soil: Recovery, resolution, and annotation of four DNA markers. *PLOS ONE*, 11(6):e0157505.

Féon, V., Schermann-Legionnet, A., Delettre, Y., Aviron, S., Billeter, R., Bugter, R., Hendrickx, F., and Burel, F. (2010). Intensification of agriculture, landscape composition and wild bee communities: A large scale study in four European countries. *Agriculture, Ecosystems and Environment*, 137(1-2):143–150.

Galan, M., Pons, J., Tournayre, O., Pierre, E., Leuchtmann, M., Pontier, D., and Charbonnel, N. (2017). Metabarcoding for the parallel identification of several hundred predators and their prey: Application to bat species diet analysis. *Molecular Ecology Resources*, 18(3):474–489.

Gallai, N., Salles, J., Settele, J., and Vassiere, B. (2009). Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. *Ecological Economics*, 68(3):810– 821.

Gill, R., Ramos-Rodriguez, O., and Raine, N. (2012). Combined pesticide exposure severely affects individual- and colony-level traits in bees. *Nature*, 491:105–108.

Godfray, H., Blacquiere, T., Field, L., Hails, R., Petrokofsky, G., Potts, S., Raine, N., Vanbergen, A., and McLean, A. (2014). A restatement of the natural science evidence base concerning neonicotinoid insecticides and insect pollinators. *Proceedings of the Royal Society B*, 281(20140558).

Goulson, D. (2013). Review: An overview of the environmental risks posed by neonicotinoid insecticides. *Journal of Applied Ecology*, 50(4).

Goulson, D., Hanley, M., Darvill, B., Ellis, J., and Knight, M. (2005). Causes of rarity in bumblebees. *Biological Conservation*, 122(1):1 – 8.

Goulson, D., Nicholls, E., Botias, C., and Rotheray, E. (2015). Bee declines driven by combined stress from parasites, pesticides, and lack of flowers. *Science*, 347(6229).

Hathaway, N., Parobek, C., Juliano, J., and Bailey, J. (2018). Seekdeep: single-base resolution *de novo* clustering for amplicon deep sequencing. *Nucleic Acids Research*, 46(4):e21.

Hawkins, J., de Vere, N., Griffith, A., Ford, C., Allainguillaume, J., Hegarty, M., and Baillie L, Adams-Groom, B. (2015). Using DNA metabarcoding to identify the floral composition of honey: A new tool for investigating honey bee foraging preferences. *PLOS ONE*, 10(8):e0134735.

Hebert, P., Cywinska, A., Ball, S., and deWaard, J. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B*, 270(1512):313–21.

Henry, M., Béguin, M., Requier, F., Rollin, O., Odoux, J., Aupine, P., Aptel, J., Tchamitchian, S., and Decourtye, A. (2012). A common pesticide decreases foraging success and survival in honey bees. *Science*, 336(6079):348–350.

Hollingsworth, P., Forrest, L., Spouge, J., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M., Cowan, R., Erickson, D., Fazekas, A., Graham, S., James, K.,

Kim, K., Kress, W. J., Schneider, H., van AlphenStahl, J., Barrett, S., van den Berg, C., Bogarin, D., Burgess, K., Cameron, K., Carine, M., Chacón, J., Clark, A., Clarkson, J., Conrad, F., Devey, D. S., Ford, C., Hedderson, T., Hollingsworth, M., Husband, B., Kelly, L., Kesanakurti, P., Kim, J., Kim, Y., Lahaye, R., Lee, H., Long, D., Madriñán, S., Maurin, O., Meusnier, I., Newmaster, S., Park, C., Percy, D., Petersen, G., Richardson, J., Salazar, G., Savolainen, V., Seberg, O., Wilkinson, M., Yi, D., and Little, D. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, 106(31):12794–12797.

Hollingsworth, P., Graham, S., and Little, D. (2011). Choosing and using a plant DNA barcode. *PLOS ONE*, 6(5):e19254.

Holovachov, O., Haenel, Q., Bourlat, S., and Jondelius, U. (2017). Taxonomy assignment approach determines the efficiency of identification of OTUs in marine nematodes. *Royal Society Open Science*, 4(8):170315.

Hung, K., Kingston, J., Albrecht, M., D.A., H., and Kohn, J. (2018). The worldwide importance of honey bees as pollinators in natural habitats. *Proceedings of the Royal Society B*, 285(1879).

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Société vaudoise des sciences naturelles*, 44:223–280.

Janda, J. and Abbott, S. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *Journal of Clinical Microbiology*, 45(9):2761–2764.

Ji, Y., Ashton, L., Pedley, S., Edwards, D., Tang, Y., Nakamura, A., Kitching, R., Dolman, P., Woodcock, P., Edwards, F., Larsen, T., Hsu, W., Benedick, S., Hamer, K., Wilcove, D., Bruce, C., Wang, X., Levi, T., Lott, M., Emerson, B., and Yu, D. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, 16(10).

Keegan, K., Glass, E., and Meyer, F. (2016). MG-RAST, a metagenomics service for analysis of microbial community structure and function. *Methods Mol Biol*, 1399:207–33.

Kienast, F., J., H., Grêt-Regamey, A., Haines-Young, R., and Potschin, M. (2019).

Landscape Planning with Ecosystem Services. Landscape Series, vol. 24. Springer, Dordrecht.

Kohler, F. (2007). From DNA taxonomy to barcoding - how a vague idea evolved into a biosystematic tool. *Zoosystematics and Evolution*, 83(S1).

Kress, W. and Erickson, D. (2007). A two-locus global DNA barcode for land plants: The coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *PLOS ONE*, 2(6):e508.

Laha, R., Mandal, S., Ralte, L., Ralte, L., Kumar, N., Gurusubramanian, G., Satishkumar, R., Mugasimangalam, R., and Kuravadi, N. (2017). Meta-barcoding in combination with palynological inference is a potent diagnostic marker for honey floral composition. *AMB Express*, 7(132).

Lejzerowicz, F., Esling, P., Pillet, L., Wilding, T., Black, K., and Pawlowski, J. (2015). High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. *Scientific Reports*, 5(Article 13932).

Li, W., Fu, L., Niu, B., Wu, S., and Wooley, J. (2012). Ultrafast clustering algorithms for metagenomic sequence analysis. *Briefings in Bioinformatics*, 13(6):656–668.

Liolios, V., Tananaki, C., Dimou, M., Kanelis, D., Goras, G., Karazafiris, E., and Thrasyvoulou, A. (2015). Ranking pollen from bee plants according to their protein contribution to honey bees. *Journal of Apicultural Research*, 54(5):582–592.

Ma, J. and Li, X. (2015). Organellar genome copy number variation and integrity during moderate maturation of roots and leaves of maize seedlings. *Current Genetics*, 61(4):591–600.

Ma, X., Shao, Y., Tian, L., Flasch, D., Mulder, H., Edmonson, M., Liu, Y., Chen, X., Newman, S., Nakitandwe, J., Li, Y., Li, B., Shen, S., Wang, Z., Shurtleff, S., Robison, L., Levy, S., Easton, J., and Zhang, J. (2019). Analysis of error profiles in deep next-generation sequencing data. *Genome Biology*, 20(1):50.

Maini, S., Pedrzycki, S., and Porrino, C. (2010). The puzzle of honey bee losses: a brief review. *Bulletin of Insectology*, 63(1):153–160.

Mallott, E., Garber, P., and Malhi, R. (2018). *trnL* outperforms *rbcL* as a DNA metabarcoding marker when compared with the observed plant component of the diet of wild white-faced capuchins (*Cebus capucinus*, primates). *PLOS ONE*, 13(6):e0199556.

Maloukh, L., Kumarappan, A., Jarrar, M., Salehi, J., El-wakil, H., and Lakshmi, T. (2017). Discriminatory power of *rbcL* barcode locus for authentication of some of United Arab Emirates (UAE) native plants. *3 Biotech*, 7(144).

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2 Part 1):209–220.

Marcel, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12.

McDonald, D., Clemente, J., Kuczynski, J., Rideout, J., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., Knight, R., , and Caporaso, J. (2012). The biological observation matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience*, 1(1):7.

Moorhouse-Gann, R., Dunn, J., de Vere, N., Goder, M., Cole, N., Hipperson, H., and Symondson, W. (2018). New universal *ITS2* primers for high-resolution herbivory analyses using DNA metabarcoding in both tropical and temperate zones. *Scientific Reports*, 8(8542 (2018)).

Moritz, C. and Cicero, C. (2004). DNA barcoding: Promise and pitfalls. *PLOS*, 2(10):e354.

Newmaster, S., Fazekas, A., and Ragupathy, S. (2006). DNA barcoding in land plants: evaluation of *rbcL* in a multigene tiered approach. *Canadian Journal of Botany*, 84(3):335–341.

Nichols, R., Vollmers, C., Newsom, L., Wang, Y., Heintzman, P., Leighton, M., Green, R., and Shapiro, B. (2018). Minimizing polymerase biases in metabarcoding. *Molecular Ecology Resources*, 18(5).

Oksanen, J., Blanchet, F., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P., O'Hara, R., Simpson, G., Solymos, P., Stevens, M., Szoecs, E., and Wagner, H. (2018). vegan: Community ecology package. R package version 2.4-6.

Peel, N., Dicks, L., Clark, M., Heavens, D., Percival-Alwyn, L., Cooper, C., Davies, R., Leggett, R., and Yu, D. (2019). Semi-quantitative characterisation of mixed pollen

samples using MinION sequencing and Reverse Metagenomics (RevMet). *Methods in Ecology and Evolution*, 10(10).

Plants For A Future, T. (2019). The Plants For A Future. Accessed: 16/01/2019.

Pompanon, F., Coissac, E., and Taberlet, P. (2011). Metabarcoding, a new way of analysing biodiversity. *Biofutur*, 319:30–32.

Potts, S., Biesmeijer, J., Kremen, C., Neumann, P., Schweiger, O., and Kunin, W. (2010). Global pollinator declines: trends, impacts and drivers. *Trends in Ecology and Evolution*, 25(6):345–353.

Pound, M., Dalgleish, A., McCoy, J., and Partington, J. (2017). Melissopalynology of honey from Ponteland, UK, shows the role of *Brassica napus* in supporting honey production in a suburban to rural settings. *Palynology*, 42(3).

Prosser, S. and Hebert, P. (2017). Rapid identification of the botanical and entomological sources of honey using DNA metabarcoding. *Food Chemistry*, 214:183–191.

Quail, M., Smith, M., Coupland, P., Otto, T., Harris, S., Connor, T., Bertoni, A., Swerdlow, H., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina Miseq sequencers. *BMC Genomics*, 13(341).

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramirez, K., Knight, C., Hollander, M., Brearley, F., Constantinides, B., Cotton, A., Creer, S., Crowther, T., Davison, J., Delgado-Baquerizo, M., Dorrepaal, E., Elliott, D., Fox, G., Griffiths, R., Hale, C., Hartman, K., Houlden, A., Jones, D., Krab, E., Maestre, F., McGuire, K., Monteux, S., Orr, C., Putten, W. v. d., Roberts, I., Robinson, D., Rocca, J., Rowntree, J., Schlaeppi, K., Shepherd, M., Singh, B., Straathof, A., Bhatnagar, J., Thion, C., Heijden, M. v. d., and de Vries, F. (2018). Detecting macroecological patterns in bacterial communities across independent studies of global soils. *Nature Microbiology*, 3(2):189–196.

Ratnasingham, S. and Hebert, P. (2007). bold: The barcode of life data system. *Molecular Ecology Notes*, 7(3):355–364.

Rauwolf, C., Golczyk, H., Greiner, S., and Herrmann, R. (2010). Variable amounts of DNA related to the size of chloroplasts iii. biochemical determinations of DNA amounts per organelle. *Molecular Genetics and Genomics*, 283(1):35–47.

Royal Horticulutural Society (2014). Royal Horticultural Society Horticultural Database. Accessed: 16/01/2019.

Schloss, P. and Westcott, S. (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for *16S* rRNA gene sequence analysis. *Applied and Environmental Microbiology*, (77):3219–3226.

Schloss, P., Westcott, S., Ryabin, T., Hall, J., Hartmann, M., Hollister, E., Lesniewski, R., Oakley, B., Parks, D., Robinson, C., Sahl, J., Stres, B., Thallinger, G., Horn, D., and Weber, C. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23):7537–7541.

Schoch, C., Seifert, K., Huhndorf, S., Robert, V., Spouge, J., Levesque, A., and Chen, W. (2012). Nuclear ribosomal internal transcribed spacer ITS region as a universal DNA barcode marker for fungi. *PNAS*, 109(16):6241–6246.

Schultz, C. and Dlugosch, K. (1999). Nectar and hostplant scarcity limit populations of an endangered Oregon butterfly. *Oecologia*, 119(2):231–238.

Selvaraj, D., Sarma, R., and Sathiskumar, R. (2008). Phylogenetic analysis of chloroplast *matK* gene from Zingiberaceae for plant DNA barcoding. *Bioinformation*, 3(1):24–27.

Sgolastra, F., Medrzycki, P., Bortolotti, L., Maini, S., Porrini, C., Simon-Delso, N., and Bosch, J. (2020). Bees and pesticide regulation: Lessons from the neonicotinoid experience. *Biological Conservation*, 241(108356).

Sickel, M., Ankenbrand, M., Grimmer, G., Holzschuh, A., Härtel, S., Lanzen, J., Steffan-Dewenter, I., and Keller, A. (2015). Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. *BMC Ecology*, 15(20).

Sipes, S. and Tepedino, V. (2005). Pollen-host specificity and evolutionary patterns of host switching in a clade of specialist bees (apoidea: Diadasia). *Biological Journal of the Linnean Society*, 86(4):487–505.

Sivakesava, S. and Irudayaraj, J. (2006). A rapid spectroscopic technique for determining honey adulteration with corn syrup. *Journal of Food Science*, 66(6).

Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., Pei, Z., Blaser, M., Aliferis, C., and Alekseyenko, A. (2013). A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome*, 1(11).

Tapolczai, K., Vasselon, V., Bouchez, A., Stenger-Kovács, C., Padisák, J., and Rimet, F. (2018). The impact of OTU sequence similarity threshold on diatom-based bioassessment: A case study of the rivers of Mayotte (France, Indian Ocean). *Ecology and Evolution*, 9(1).

Topitzhofer, E., Lucas, H., Chakrabarti, P., Breece, C., Bryant, V., and Sagili, R. (2019). Assessment of pollen diversity available to honey bees (hymenoptera: Apidae) in major cropping systems during pollination in the wester united states. *Journal of Economic Entomology*, 112(5):2040–2048.

Valentini, A., Miquel, C., and Taberlet, P. (2010). DNA barcoding for honey biodiversity. *Diversity*, 2:610–617.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P., Bellemain, E., Besnard, A., Coissac, E., F., B., Gaboriaud, C., Jean, P., Poulet, N., N., R., Copp, G., Geniez, P., Pont, D., Argillier, C., Baudoin, J., Peroux, T., Crivelli, A., Olivier, A., Acqueberge, M., Le Brun, M., Møller, P., Willerslev, E., and Dejean, T. (2015). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, 25(4).

Veltri, K., Espiritu, M., and Singh, G. (1990). Distinct genomic copy number in mitochondria of different mammalian organs. *Journal of Cellular Physiology*, 143(1):160–164.

Von Der Ohe, W., Persano Oddo, L., Piana, M., Morlot, M., and Martin, P. (2004). Harmonized methods of melissopalynology. *Apidologie*, 35:S18–S25.

Yoccoz, N. G., Brathen, K. A., Gielly, L., Haile, J., Edwards, M. E., Goslar, T., Von Stedingk, H., Brysting, A. K., Coissac, E., Pompanon, F., Sonstebo, J. H., Miquel, C., Valentini, A., De Bello, F., Chave, J., Tuiller, W., Wincker, P., Cruaud, C., Gavory, F., Rasmussen, M., Gilbert, M. T. P., Orlando, L., Brochmann, C., Willerslev, E., and

Tabelet, P. (2012). DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology*, 21(15):3647–3655.

Yu, J., Xue, J., and Zhou, S. (2011). New universal *matK* primers for DNA barcoding angiosperms. *Journal of Systematics and Evolution*, 49(3).

Zimmerman, J., Glockner, G., Jahn, R., Enke, N., and Gemeinholzer, B. (2014). Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. *Molecular Ecology Resources*, 15(3):526–42.

5.9 Tables and Figures

Table 5.1

Details of honey samples. Site (Postcode): the locations of the 15 bee hives sampled, Quantity (g): the amount of honey which was used for DNA extraction (40g where available), Conc. (ng/µL): concentration of purified DNA, Sample Type: descriptions of the honey material sampled from each hive.

Table 5.1

#	Site (Postcode)	Quantity (g)	Conc. (ng/µL)	Sample Type
1	The Printworks, Manchester (M4 2BS)	40.00	1165.7	Honey
2	Manchester Cathedral, Manchester (M3 1SX)	40.00	806.2	Honey
3	Chorlton Meadow, Manchester (M21 9ET)	40.00	607.5	Honey
4	Warrington, Cheshire (WA5 0DJ)	40.00	452.5	Honey
5	Manchester Art Gallery, Manchester (M2 3JL)	40.00	1534.9	Honey
6	Manchester Museum, Manchester (M13 9PL)	40.00	521.1	Honey
7	Heaton Park, Manchester (M25 2SW)	40.00	732.7	Honey
8	Chorlton, Manchester (M21 9DF)	40.00	1191.5	Chunk honey
9	Ashton-under-Lyne, Manchester (OL7 9NY)	1.51	33.7	Unfiltered
10	Northenden, Manchester (M22 4WS)	5.96	236.3	Honey
11	Swinton, Manchester (M27 4FY)	40.00	670.3	Honey
12	Rossendale, Lancashire (BB4 7DQ)	28.23	182.7	Honey
13	Bolton, Greater Manchester (BL2 5BF)	9.82	30.8	Honey
14	Stockport, Greater Manchester (SK6 1EJ)	15.46	358.4	Honey
15	Ashton, Manchester (M33 5PE)	30.00	156.8	Comb

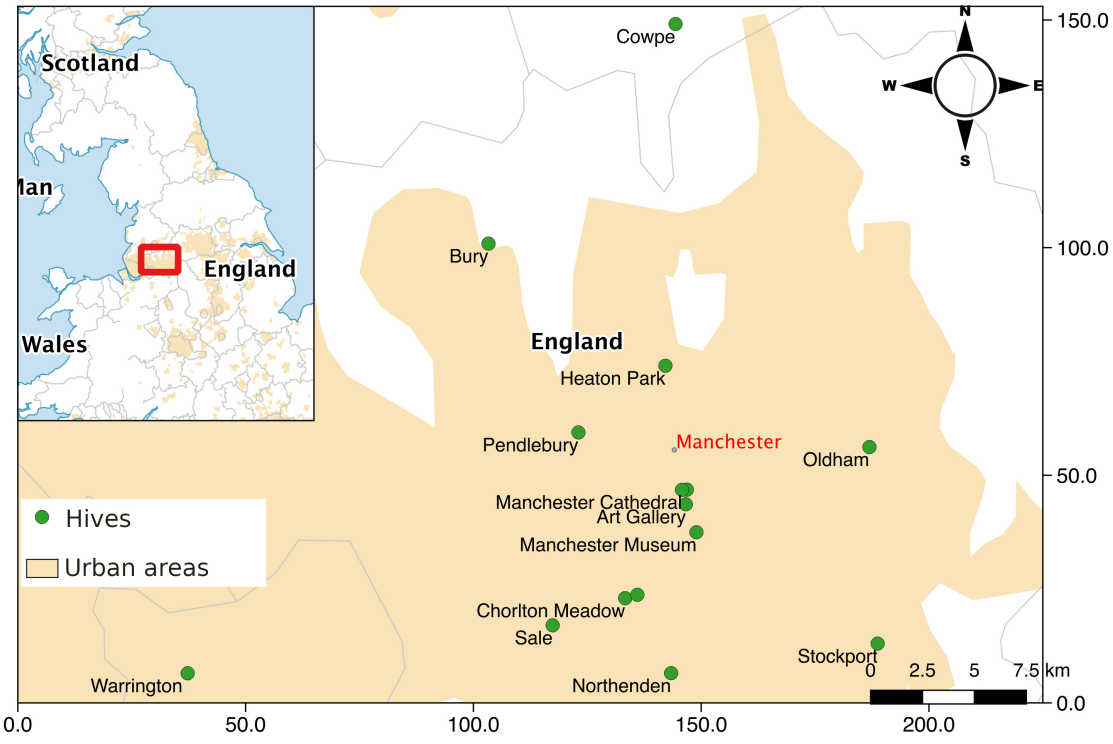

Scotland
England
Wales
Cowpe
Bury
England
Heaton Park
Pendlebury
Manchester
Oldham
Manchester Cathedral
Art Gallery
Manchester Museum
Chorlton Meadow
Sale
Stockport
Warrington
Northenden
Hives
Urban areas
N
W
E
S
0
2.5
5
7.5 km
150.0
100.0
50.0
0.0
0.0
50.0
100.0
150.0
200.0

Figure 5.1

Table 5.2

Primer	Marker	Primer Sequence (5′- 3′)	Amplicon Size (bp)
rbcLa-F[9]	rbcL (forward)	ATGTCACCACAAACAGAGACTAAAGC	500-600
rbcLR590	rbcL (reverse)	AGTCCACCGCGTAGACATTCAT	
S2F[10]	ITS2 plant (forward)	ATGCGATACTTGGTGTGAAT	450-550
S3R	ITS2 plant (reverse)	GACGCTTCTCCAGACTACAAT	
ITS2-u3-F[11]_F	ITS2 universal (forward)	CAWCGATGAAGAACGYAGC	450-500
ITS2-u4-R	ITS2 universal (reverse)	RGTTTCTTTTCCTCCGCTTA	

Table 5.3

Assignment Method	rbcL	ITS2u	ITS2p	Mean (+/- 95% CI)
UCLUST	74.8%	76.5%	12.6%	55% (45.2% - 64.4%)
BLASTn	25.1%	21.9%	87.1%	45% (35.6% - 54.8%)

Table 5.4

Species	*rbcL*	*ITS2u*	
rbcL		0.001 (0.601)	
ITS2u	0.006 (0.756)		
Genus	*rbcL*	*ITS2u*	*ITS2p*
rbcL		0.002 (0.612)	0.100 (0.011)
ITS2u	0.008 (0.794)		0.153 (0.005)
ITS2p	0.048 (0.035)	0.112 (0.01)	
Family	*rbcL*	*ITS2u*	*ITS2p*
rbcL		0.010 (0.213)	0.115 (0.019)
ITS2u	0.002 (0.323)		0.143 (0.004)
ITS2p	0.061 (0.035)	0.098 (0.014)	
Order	*rbcL*	*ITS2u*	*ITS2p*
rbcL		0.001 (0.428)	0.088 (0.015)
ITS2u	0.000 (0.528)		0.101 (0.006)
ITS2p	0.067 (0.012)	0.074 (0.012)	


Average Sequence / OTU counts after each stage of data processing
log2 Counts
2^20
2^15
2^10
2^5
Raw Sequences
Cutadapt
Chimera Removal
OTUs 97%
Unique OTUs
Genus in UK
Species in UK
Stage of Data Processing
Legend
ITS_plant
ITS2_universal
rbcL

Figure 5.2

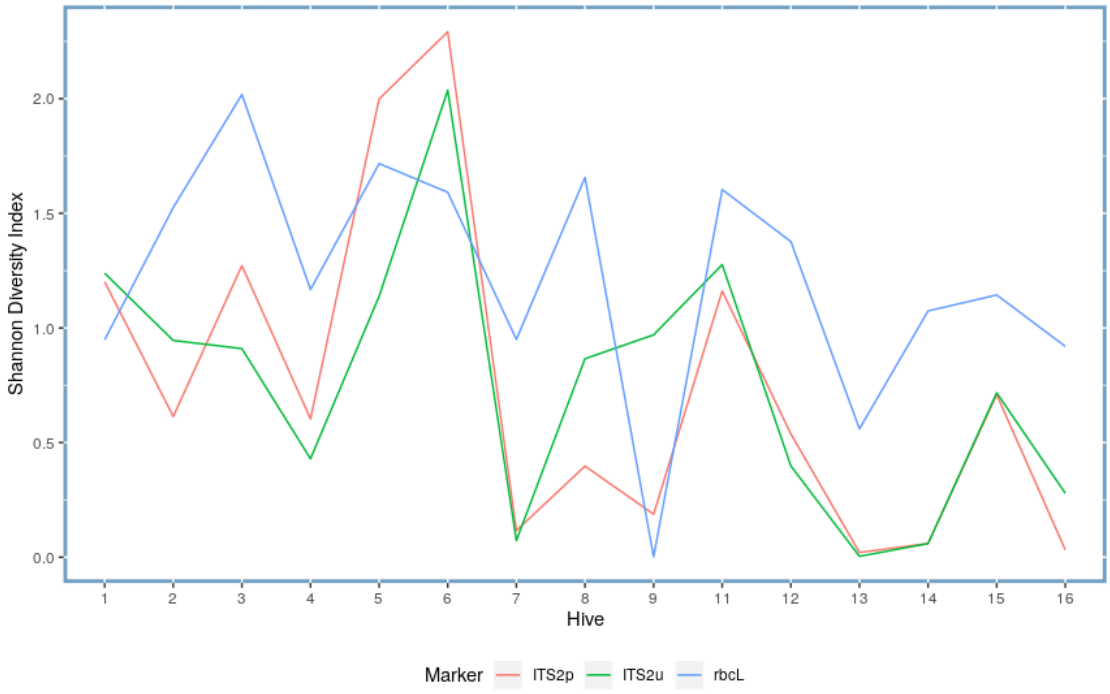

Shannon Diversity Index
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
8
9
11
12
13
14
15
16
Hive
Marker
ITS2p
ITS2u
rbcL

Figure 5.3

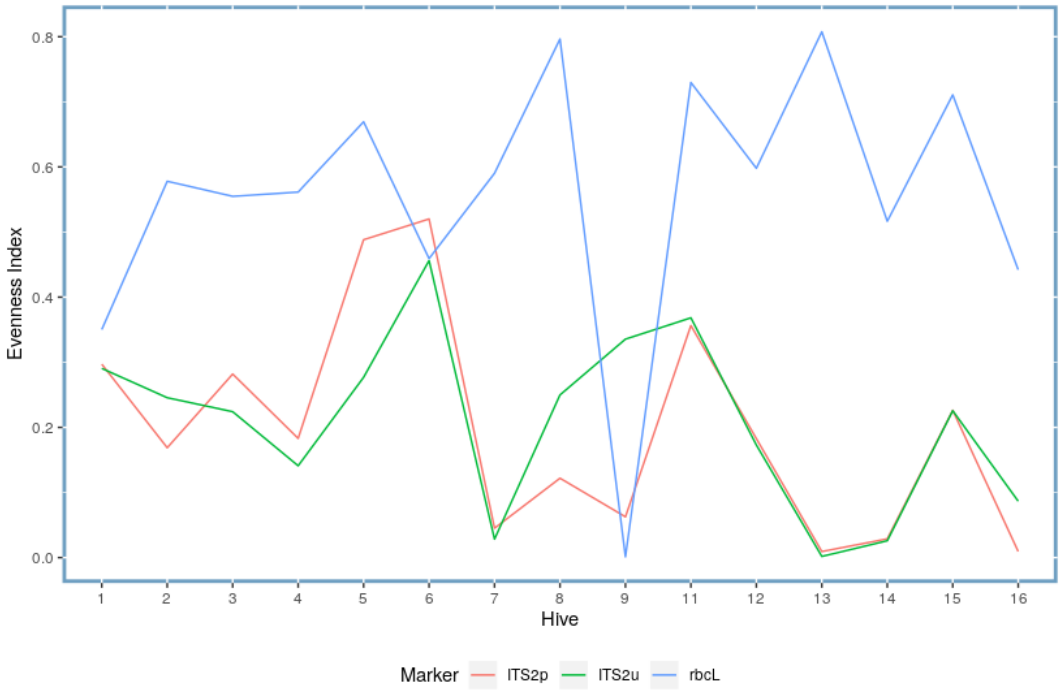

0.8
0.6
0.4
0.2
0.0
Evenness Index
1
2
3
4
5
6
7
8
9
11
12
13
14
15
16
Hive
Marker
ITS2p
ITS2u
rbcL

Figure 5.4

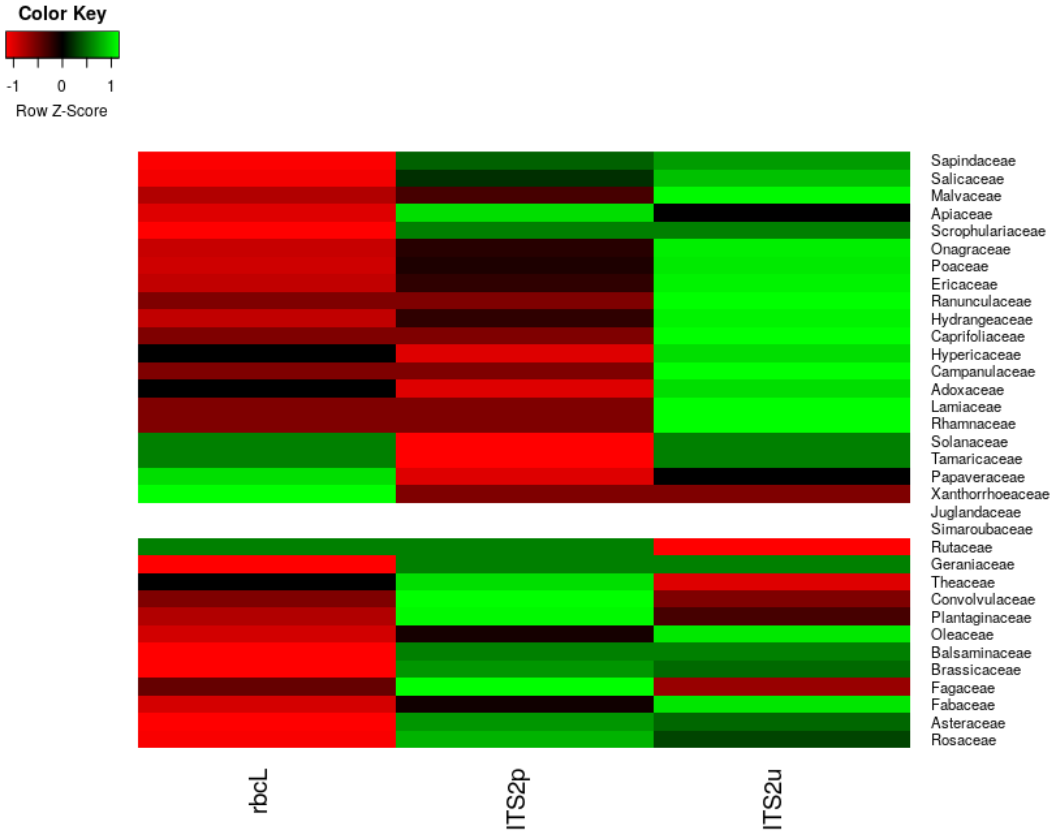

Color Key
-1
0
1
Row Z-Score
Sapindaceae
Salicaceae
Malvaceae
Apiaceae
Scrophulariaceae
Onagraceae
Poaceae
Ericaceae
Ranunculaceae
Hydrangeaceae
Caprifoliaceae
Hypericaceae
Campanulaceae
Adoxaceae
Lamiaceae
Rhamnaceae
Solanaceae
Tamaricaceae
Papaveraceae
Xanthorrhoeaceae
Juglandaceae
Simaroubaceae
Rutaceae
Geraniaceae
Theaceae
Convolvulaceae
Plantaginaceae
Oleaceae
Balsaminaceae
Brassicaceae
Fagaceae
Fabaceae
Asteraceae
Rosaceae
rbcL
ITS2p
ITS2u

Figure 5.5

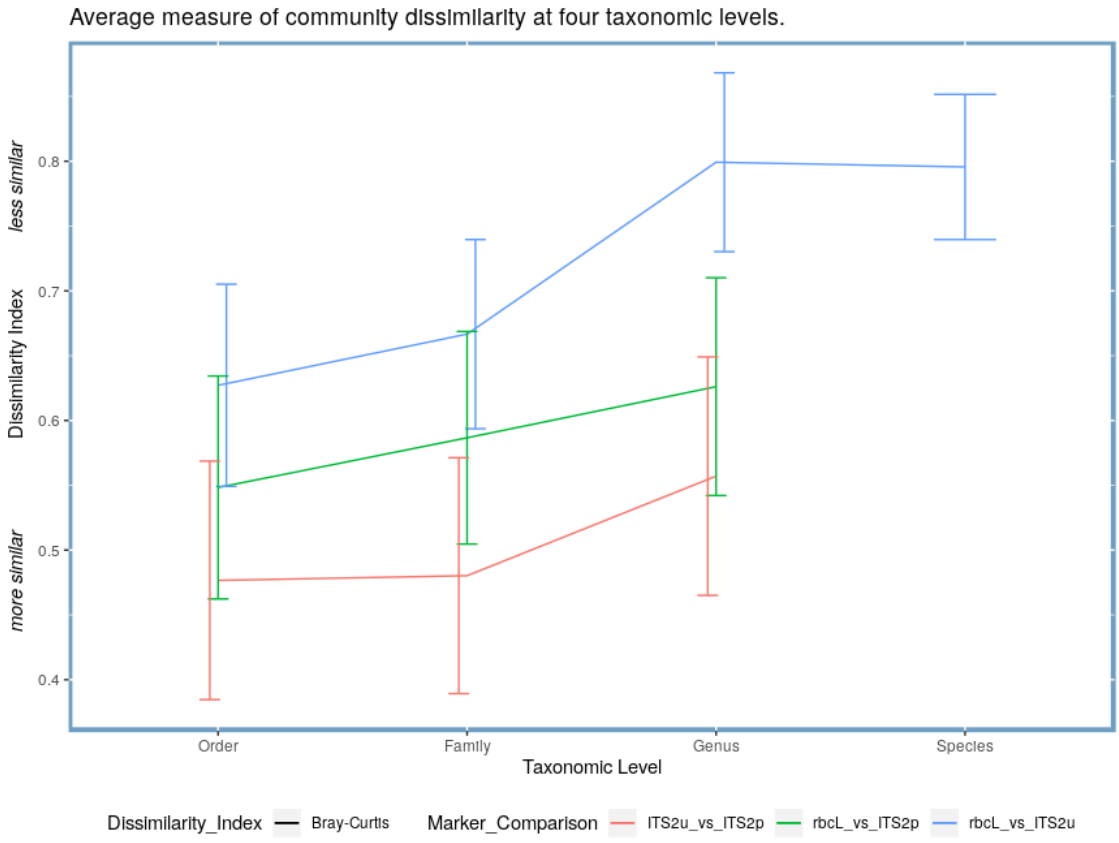

Average measure of community dissimilarity at four taxonomic levels.
less similar
more similar
Dissimilarity Index
0.8
0.7
0.6
0.5
0.4
Order
Family
Genus
Species
Taxonomic Level
Dissimilarity_Index
Bray-Curtis
Marker_Comparison
ITS2u_vs_ITS2p
rbcL_vs_ITS2p
rbcL_vs_ITS2u

Figure 5.6

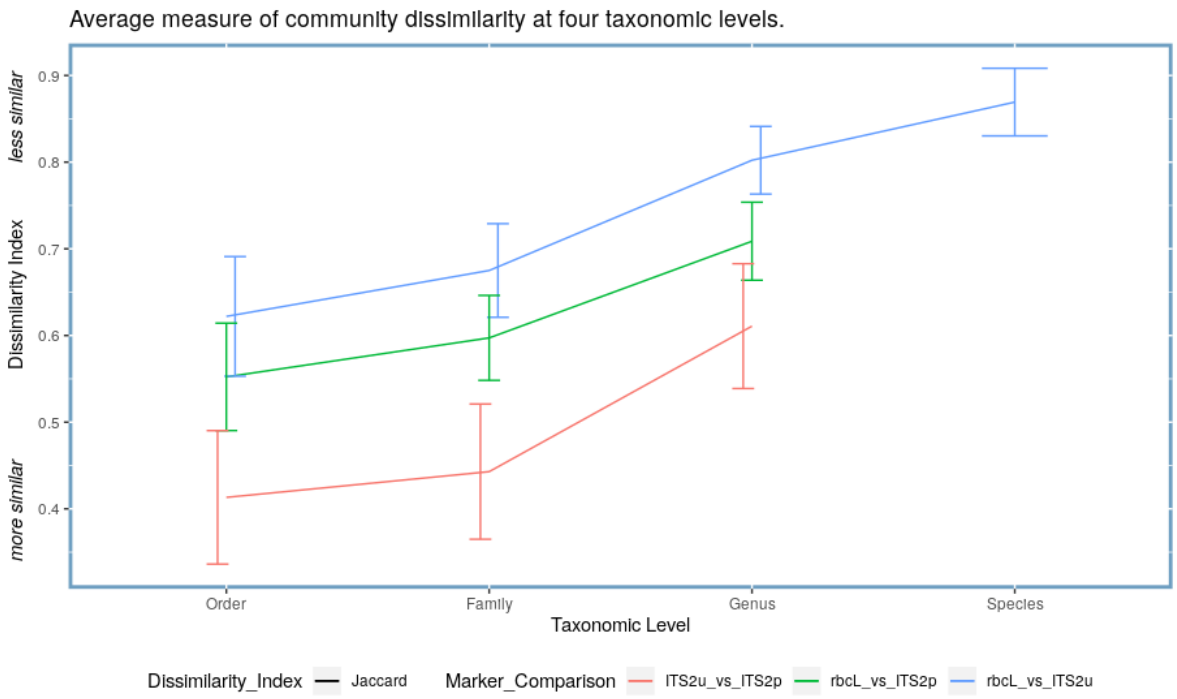

Average measure of community dissimilarity at four taxonomic levels.
less similar
more similar
Dissimilarity Index
0.9
0.8
0.7
0.6
0.5
0.4
Order
Family
Genus
Species
Taxonomic Level
Dissimilarity_Index
Jaccard
Marker_Comparison
ITS2u_vs_ITS2p
rbcL_vs_ITS2p
rbcL_vs_ITS2u

Figure 5.7

Chapter 6
General Discussion.

6.1 General Discussion

Across the globe, ecosystems are being damaged, in many cases irreparably, by the actions of humans through population increase and the associated consumption of resources. Effects including the loss of ecosystem services, destruction of available habitat and reduction of biodiversity have been recorded in every type of ecosystem whether marine, terrestrial, montane, arboreal or any other ecosystem classification (Frankham et al., 2004; Sodhi and Ehrlich, 2010). In 2019 alone, we have recorded unprecedented fires in the arctic and the Amazon rainforest, had the outlook for the Great Barrier Reef downgraded to very poor, and in the UK recorded the hottest ever temperature and second wettest summer, all of which can be directly linked to the actions of humans (Vautard et al., 2019; Escobar, 2019; Ley, 2019).

The understanding of the functioning of species and ecosystems is critical to their management, conservation and restoration. The development of molecular forms of analysis, from marker based studies to the development of true genomic approaches contribute valuable information towards these goals. This thesis presents research intended to exhibit the value and power of molecular analysis methods to answer important ecological questions, and also to demonstrate some limitations which are equally important. Since the development of next-generation sequencing (NGS) technologies, a revolution has occurred in the scale and availability of nucleotide sequence data with almost no area of biological sciences untouched by the developments. For questions of ecology in particular, this revolution in the availability of sequence data has been particularly powerful.

Chapter two involved performing shotgun, whole genomic sequencing (WGS) of the genome of the endangered skate (*Raja undulata*) and the subsequent development and characterisation of a novel panel of microsatellite markers. I used Illumina Nextera sequencing and a bioinformatics pipeline, to design a panel of microsatellite markers (Griffiths et al., 2016). This was the first time that WGS had been performed, and microsatellite markers had been developed in this species. Once developed these PCR markers were used to assess a population of the species spread throughout an aquaria network in the UK, which were being actively managed. The data generated in this chapter was used to inform the breeding strategy employed by the aquaria, and the markers are available for assessment of future generations, and other closely related species. Calculations of linkage disequilibrium (LD) in the captive animals revealed an

interesting characteristic. Across the whole captive population, 48% of pairs of microsatellite markers showed significant rates of LD, but no pair of markers had significant values when just the wild-caught individuals were analysed. An admixture event in a populations recent history is known to cause an associated increase in LD (Slate and Pemberton, 2007) due to animals being recent descendants of common ancestors. As wild animals were caught from different regions of the UK, we can assume relatively low relatedness between them, and therefore the establishment of the captive population likely represents an admixture event. This effect upon rate of LD will likely decrease in future generations. I also recorded high rates of heterozygosity in the captive population, both in the wild-caught and the captive bred individuals, with no significant difference between either cohort. This result indicates a maintenance of high allelic diversity into this first generation of captive individuals. I established that in this early stage of the management of the captive population (there had only been a single captive generation at the time the study was performed), that there were no detectable signs of any negative genetic effects associated with the small population size. Furthermore, I made the case for the inclusion of genetic data in the management of any small population such as this, and explain the benefits and limitations of genetic markers of this type. I have shown the feasibility and benefit of generating a panel of bespoke microsatellite markers for the analysis of a population. The methods employed, including the bioinformatics methods to mine sequence data for microsatellite regions, and to design and optimise PCR markers are extremely useful wherever non-model species require markers. I demonstrated a now well-documented molecular and bioinformatics workflow for marker development which is applicable to any species of conservation concern.

The development of new microsatellite markers for use with a species is extremely common, and microsatellites remain an extremely popular and powerful marker for many types of analysis (Vieira et al., 2016; Ribout et al., 2019). In chapter three I used NGS data to develop a new bioinformatics approach, and method by which microsatellite markers can be developed in species. This is of particular importance to non-model species where there is currently not sequence data, or an assembled genome available. The novel method allows researchers to choose the optimum potential microsatellite markers by allowing the *in silico* detection of polymorphic loci, loci with high levels of PCR primer conservation and the avoidance of markers likely to result in null alleles, or those that break assumptions of the microsatellite evolution model. This novel method replaces several of the laboratory based

quality control processes which much be performed on new markers with an automated computer program, and has the potential to streamline the design process. My results demonstrate that the application of this new method, results in the filtering of a list of potential markers in the tens of thousands, down to a list of hundreds of high quality markers. The development of these markers can then continue with significantly increased rate of success, reducing wasted investment in failed markers and streamlining the entire microsatellite marker development process. This process makes the production of a panel of markers quicker and cheaper, and therefore more widely available to ecologists. The new method has already been used to develop several novel microsatellite marker panels in our lab group and is now available for ecologists everywhere to use in their marker design workflows.

The fourth chapter contains a study of the population structure of the European lobster (*Homarus gammarus*) around the coasts of the UK and Ireland. I used microsatellite markers to genotype samples from 15 sites, and performed statistical analysis to measure gene flow and connectivity. In a parallel study, I used a restriction site associated DNA sequencing methodology (RAD-Seq) to develop a panel of SNP markers, which were also used for an analysis of the genetic population structure. Using two types of genetic marker allowed me to conclude that there is very little genetic structure in *H. gammarus* at a national scale around the UK and Ireland. This is an important result for the management strategy of the species, which is at risk of over-exploitation due to the high value associated with lobster. Our results have been transmitted to the various conservation authorities managing the fishing stocks and will be used to help inform future conservation. The marker comparison was the first direct comparison in the statistical power and ease of application of microsatellite and SNPs in a larval dispersing organism. This work contributes to the current debate regarding the suitability of different types of molecular marker for population genetics, and is useful to researchers planning such a study.

As well as investigating intra-species variation, such as the pedigree analysis and population genetics I have discussed previously, NGS is also an extremely powerful tool for species identification. It is particularly effective in the analysis of mixed, complex or cryptic communities. In chapter five I investigated some of the biases inherent in metabarcoding analysis, with specific reference to the analysis of the plant community detected in pollen from the honey of *Apis mellifera* hives from Greater Manchester, UK. We are increasingly aware that every stage of a standard metabarcoding workflow

increases biases into the results (D'Amore et al., 2016). One of the major components of the library preparation is the choice of barcoding gene which is chosen, as it is variation at this particular region of the genome taken to be representative of the wider community. In this chapter I performed parallel analyses of the plant communities using three barcoding markers and assess the variability between results, highlighting the caution which researchers should have whilst interpreting results. This then allowed me to make recommendations regarding the analysis of metabarcoding data generally, the importance of using multiple markers with specific reference to the analysis of a mixed plant community, and the importance of using a high quality reference database. There are an abundance of tools and databases available for metabarcoding of microbial communities, however the resources are not as developed for research into mixed plant communities. The methods developed, and results described in this chapter are intended to combat some of these limitations, and to progress the confidence in plant metabarcoding.

6.2 Thesis Achievements

The aim of the thesis was to use DNA analysis methods alongside computation techniques to answer several ecological questions. An assessment as to how well each of the goals stated in section 1.5 follows:

(A) I was able to successfully use next-generation sequencing, and bioinformatics, to develop and optimise a panel of microsatellite markers. I subsequently used these markers to perform an assessment of the genetic health of a population of *R. undulata*. Using limited markers I was able to demonstrate the utility of developing novel markers and rapidly applying them to a population of conservation interest. I recorded no evidence of negative genetic effects as a result of the small population size, and provide PCR primers and conditions for further assessments and population genetic studies of the species to further aid their conservation.

(B) The development of novel microsatellite markers is now a very common process where previously published markers are not available. In this study I was able to conceive a novel approach to bioinformatics driven microsatellite marker design; one which streamlined the laboratory testing of markers by performing as much quality

control *in silico* as possible. I designed and programmed a new bioinformatics tool, and demonstrated the utility in two species.

(C) Using genetic markers (both microsatellites and single-nucleotide polymorphisms) I was able to genotype several hundred wild *H. gammarus* samples, from sites around the coast of the UK and Ireland. Both parallel studies gave a congruent result: that of an almost complete absence of detectable population structure in the species, in the region. This result is in keeping with other similar studies and is informative to the management of the fisheries by the conservation authorities responsible.

(D) After the parallel analysis of wild *H. gammarus* samples using microsatellite and single-nucleotide polymorphisms (SNPs), I was able to show that in this instance, both marker types gave the same result. The benefit of SNPs over microsatellites is one of ease and time of application, whilst the overall cost was similar. Given that neither marker was able to provide more statistical power to describe structure than the other, in this instance our comparative study suggests that SNPs would be the preferable marker type for population genetic analysis of a larval dispersing marine organism, such as *H. gammarus*.

(E) I successfully used next-generation sequencing and bioinformatics methods to implement a metabarcoding experiment upon DNA extracted from pollen in honey. By analysing three different plant metabarcoding markers in parallel I investigated variation in the resulting community descriptions, considering the generation of false positive results, the limits to their taxonomic resolution and the potential power of using multiple molecular barcodes to increase the confidence in taxonomic assignments.

6.3 Evaluation of Methods

The implementation and assessment of methods have formed an integral part of the research in this thesis. As well as providing answers to ecological questions, I have demonstrated the power, and pitfalls of just some of the wide-range of molecular tools available to ecologists. The work presented across the four data chapters use several

technologies and methodologies.

6.3.1 Application and Analysis of Microsatellite Markers

In molecular ecology and conservation, microsatellites are known primarily for their function as useful genetic markers for pedigree analysis and population genetics, for example. However, in the genome of the species, microsatellites perform, or can be involved in, a range of important processes, can be situated in introns or exons, and may be under strong selective pressure due to functional importance (Li et al., 2004). Microsatellites are known to be very highly abundant throughout the genome, but are non-randomly distributed, suggesting an element of selective pressure not always in keeping with theoretical models of their neutral evolution (Tautz and Renz, 1984). Rather, microsatellites are now known to be functionally important in gene transcription, translation, recombination, DNA replication, amongst many other important processes (Treco and Arnheim, 1986; Dutreix, 1997). As a marker, their power comes from the highly polymorphic nature of the number of repeats found between individuals. In some instances, the repeat number of a microsatellite associated with a particular gene is a key factor in regulating gene expression and expression levels, and can even act as an on/off switch (Liu et al., 2000). Gene associated markers, such as gene associated single-nucleotide polymorphisms (SNPs), can also be used to identify loci which are highly divergent between target groups, and these are often markers which do not conform to neutral genetic models, but instead are under strong selective pressure. Genome scans to identify hundreds of thousands of these powerful markers can be used to provide extremely powerful population assignment methods, however these must be targeted specifically using design methods or mined from random DNA sequencing methods, such as RAD-Seq (Nielsen et al., 2012).

I used microsatellite markers for analysis in chapters two and four. The markers were used to measure relatedness between individuals and inform management of a small population of rays (*Raja undulata*), and to measure genetic connectivity between wild populations of the European lobster (*Homarus gammarus*) Broadly, the application of PCR primers to amplify microsatellites, capillary electrophoresis and analysis using peak scoring software was extremely successful. I was able to generate several PCR multiplexes using a three primer method to add fluorescent dyes onto PCR amplicons, which functioned very efficiently (Blacket et al., 2012). There were some minor setbacks when analysing genotype data derived from the markers. In chapter two, only eight

microsatellite markers were analysed, which limited the extent of my statistical power to make inferences regarding the population. One marker was subsequently removed from analysis due to exhibiting significant deviation in allele frequencies from those expected under Hardy-Weinberg equilibrium (HWE). The population from which my samples were taken did not adhere to several of the assumptions of Hardy-Weinberg, mainly that I was not sampling from a large, unrelated population. An analysis of these samples using many more microsatellite markers (or SNPs) would likely give greater statistical power and would allow accurate pedigrees to be established as parent-progeny assignments could be made. Unfortunately, I did not have sufficient genetic data in this chapter to confidently generate these pedigrees.

In chapter four, two markers (out of 20 amplified) ultimately had to be removed from the *H. gammarus* data set due to inconsistent rates of amplification. I decided upon a threshold where a marker must successfully amplify, and produce a clearly interpretable trace, in 66.6% of samples for the data to be used in further analysis. Error in interpreting microsatellites traces, or resulting from sample contamination does not appear to be an issue, as the calculated error rate was low at 1.74%. I saw a low frequency of marker/sites where a microsatellite marker was estimated to contain relatively high rates of null alleles. Given the overall extremely high levels of genetic connectivity we saw between all sites, I determined that null alleles were not a significant concern and did not impact upon the validity of our results.

6.3.2 Illumina Next-Generation Sequencing Overview

It has been stated *ad infinitum* that the invention of next-generation sequencing technologies was an epoch defining moment for the biological sciences, and therefore also ecology. As it was refined over the last decade, the Illumina sequencing platforms have emerged as the dominant force in high-throughput sequencing, I used two sequencing instruments from the Illumina range for the NGS data generation in the chapters of this thesis.

6.3.3 Illumina MiSeq Sequencing

Illumina MiSeq sequencing, using the Nextera library preparation chemistry was used in chapters two and three for the shotgun genomic sequencing required for microsatellite marker design. This approach is a relatively cheap, method of generating a large quantity

of non-targeted genomic sequence data. By digesting genomic DNA (gDNA) with the transposase enzyme, adding sequencing adapters and indexes, and sequencing, it is a fast, cost effective method of data generation. In these chapters, the resulting nucleotide data sets were mined for microsatellite regions using bioinformatics software (Castoe et al., 2015; Fox et al., 2019) and novel microsatellite markers designed in several species. We were able to consistently generate very high quality sequence data in every instance. A standard quality control workflow was applied to each data set which removed any low quality regions (Bolger et al., 2014), however this resulted in very few sequence removals in any case.

6.3.4 Illumina NextSeq Sequencing

Restriction site associated DNA sequencing (RAD-Seq), commonly used for the generation and genotyping of SNP markers, typically requires a greater depth of sequence coverage than for microsatellite discovery, and as such a sequencer with a greater capacity is required. In chapter four, I performed RAD-Seq analysis upon 95 gDNA samples using the Illumina NextSeq sequencer. Based upon the same fundamental Illumina chemistry as the MiSeq, the NextSeq enables much greater depth of sequencing coverage through the significantly greater number of sequencing reads generated. The trade off however occurs in that one must sacrifice sequencing read length for greater numbers of raw reads. I used the 2* 150bp, paired end sequencing methodology to sequence gDNA which had been sonicated, and digested with the SbfI restriction enzyme. The result of these two procedures is a library of random fragment of gDNA, but all of which begin from a location in the genome matching the cutsite of the enzyme (TGCA-GG, in this case). Fragments not starting with this particular sequence are removed during the library preparation. The overall outcome of generating a sequencing library in this manner is that the random nature of shotgun sequencing is retained, but covering much less of the genome, at a greater sequencing depth. This, and similar methods are described as producing reduced coverage genomic sequence data sets. The sequencing performed well, producing high quality sequence data but I did receive less sequence data than I anticipated during the experimental design. Diagnostic analysis by Illumina technical support determined that this was a result of an under-clustered flowcell. Sequence coverage is very influenced by the size of data generation during sequencing, and as the flowcell under-performed, this reduced the sequence coverage I was able to achieve. During the SNP analysis of the RAD-Seq data, potential marker

loci are filtered by the proportion of samples in which they have been sequenced. A typical threshold is to use markers which are present in 80% of samples (Paris et al., 2017), however I could not achieve this threshold. Instead I used a threshold of 40%, which is low for an SNP based approach to population genetics. During statistical analysis, I found highly congruent results from my SNP data set, and also my microsatellite data set which was of a much higher quality, indicating that in this instance, low coverage of SNPs does not appear to have negatively impacted the results.

6.4 Thesis Conclusions and Future Direction

In this thesis, I have established the benefits that inclusion of genetic data can bring to a range of ecological questions. I generated new microsatellite markers for *Raja undulata*, used them to genotype a captive population and provided information on their relatedness to help their future conservation. I developed a novel method of detecting high quality microsatellite markers from sequence data, reducing expensive laboratory testing of markers. This new technique brings bespoke marker development within the scope of more laboratories and conservationists. I have used genetic marker to assess the population structure of *Homarus gammarus* fisheries around the UK and Ireland and this information will be used to conservation authorities to inform their management of the fisheries. Further, I performed a comparative analysis of microsatellite and SNPs, the first in a larval dispersing decapod, to compare their effectiveness for a study of this type. This research will be highly informative for other conservationists in this field planning a population genetics study. Finally, I performed metabarcoding analysis to determine the plant forage of *Apis mellifera* hives in Greater Manchester, and investigated some metabarcoding biases influenced by genetic marker choice. This has important connotations for further research into plant metabarcoding, and adds to the wider debate about how best to perform metabarcoding experiments.

There are a multitude of ways in which this wide-ranging research could be expanded upon. Discussing NGS and bioinformatics is to discuss technical innovation of sequencing hardware, as well as the interpretation of the biological results. It will therefore not be a surprise that I suggest that the next great developments in molecular ecology will come from the widespread adoption of the third-generation of DNA sequencing. Single-molecule sequencing promises to take molecular approaches past the age of PCR, into the analysis of individual strands of DNA. This single molecule resolution, and ever increasing data

throughput of sequencers, will likely lead us to a generation of molecular biologists whom do not consider genetic markers in the same way that we consider them today. Whole genome approaches will be within reach for any field of analysis. However, in the immediate future, with practice already struggling to keep up with the best current research methods, I see more incremental improvements in our marker-based analyses (Gossa et al., 2015; Sunderland et al., 2009).

Third generation sequencing technologies do not require the bioinformatic assembly of short reads, nor suffer from the PCR bias introduced by clonal amplification. New methods and techniques enabled by technological innovations, such as whole genome amplification (WGA), allow DNA barcoding approaches to be used on mixed, complex communities but without the additional complications, and influences which PCR amplification using primers can have on results (Pinar et al., 2020). Metatranscriptomics is another promising, new methodology being developed based upon the ability of the MinION Nanopore sequencing platforms ability to sequence RNA directly, as opposed to requiring reverse transcription and PCR as in previous iterations of sequencing technologies. Metatranscriptomics of environmental samples are capable of species detection, associated with more commonplace metabarcoding approaches, but also functional analysis and it has been demonstrated that mRNA transcripts from metatranscriptomic analysis have been associated closely with genes for essential metabolic processes (Semmouri et al., 2020). Whole, large genome assembly by nucleotide sequencing has been possible since the days of the human genome project (Venter et al., 1998), however now the assembly of complete genomes can be performed using a single sequencer run. Ultra-long reads generated by platforms such as the MinION Nanopore enable large fragments of chromosome to be sequenced intact, resulting in no requirement for intensive construction of contigs, and providing high quality reference constructs against which deeper sequencing data can be assembled into high confidence genome level sequences (Wang et al., 2020). These are just a few examples of the great technological innovations which are already being made as long-read, third generation sequencing platforms become more widely available, have more laboratory testing, and new bioinformatics methods are developed.

Whilst I do envisage a time when microsatellite markers will be completely superseded by SNP, or other types of marker, I do not see this happening soon. As such I can see scope to further improve upon the *in silico* marker design methods I have developed here, in chapter three. It should be possible to detect and measure linkage disequilibrium,

and calculate allele frequencies (and therefore deviation from HWE etc.), in sequence data given sufficient amounts of data generated. If several orders of magnitude more data were generated, more genomes could be included in the design process, and increased coverage mean that loci were sequenced in most, or all, individuals. The MiMi tool already derives microsatellite genotypes, based upon the number of nucleotides between the primer regions, and there is no reason why can this functionality cannot be expanded upon to perform comparisons between genotypes, such as those required for LD calculations and the like. This would serve to further automate the microsatellite design process, reducing marker design costs, and improving the availability of these important markers.

I have discussed at length some of the inherent biases involved in metabarcoding analysis. One of the fundamental issues causing these biases in the use of genetic barcoding markers, with variable, or limited resolution to discriminate between closely related taxa. The current generation of NGS platforms, have very relatively short read-lengths, compared even to older sequencing methods such as Sanger sequencing and Roche 454 pyrosequencing. With the newer platforms becoming available, such as those by Oxford Nanopore, the limitations of read-length are completely removed. The analysis of significantly longer sections of barcoding gene, will give much greater power to resolve species, but will however require the associated reference sequences and reference databases to become available. Ultimately, tools for genome-wide taxa identification, which do not require PCR amplification, and the the use of sequencers which are not read-length limited is the likely destination. By removing the biasing step of the selection and amplification of a barcoding marker, we will likely be able to remove one of the most influencing factors in metabarcoding analysis, as we move towards complete metagenomics.

References

Blacket, M., Robin, C., Good, R., Lee, S., and Miller, A. (2012). Universal primers for fluorescent labelling of PCR fragments–an efficient and cost effective approach to genotyping by fluorescence. *Molecular Ecology Resources*, 12(3):456–63.

Bolger, A., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–20.

Castoe, T., Poole, A., de Koning, A., Jones, K., Tomback, D., Oyler-McCance, S., Fike, J., Lance, S., Streicher, J., Smith, E., and Pollock, D. (2015). Correction: Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PLOS ONE*, 7(2):e30953.

D'Amore, R., U.Z., I., Schirmer, M., Kenny, J., Gregory, R., Darby, A., Shakya, M., Podar, M., Quince, C., and Hall, N. (2016). A comprehensive benchmarking study of protocols and sequencing platforms for *16S* rRNA community profiling. *BMC Genomics*, 17(55).

Dutreix, M. (1997). (GT)n repetitive tracts affect several stages of RecA-promoted recombination. *Journal of Molecular Biology*, 273(1):105–13.

Escobar, H. (2019). Amazon fires clearly linked to deforestation, scientists say. *Science*, 365(6456):853.

Fox, G., Preziosi, R., Antwis, R., Benavides-Serrato, M., Combe, F., Harris, W., Hartley, I., de Kort, S., Nekaris, A., and Rowntree, J. (2019). Multi-individual Microsatellite identification: a multiple genome approach to microsatellite design (MiMi). *Molecular Ecology Resources*, 19(6):1672–1680.

Frankham, R., Ballou, J., and Briscoe, D. (2004). *A Primer of Conservation Genetics.* Cambridge University Press, Cambridge, USA.

Gossa, C., Fisher, M., and Milner-Gulland, E. (2015). The research-implementation gap: how practitioners and researchers from developing countries perceive the role of peer-reviewed literature in conservation science. *Oryx*, 49(1):80–87.

Griffiths, S., Fox, G., Briggs, P., Donaldson, I., Hood, S., Richardson, P., Leaver, G., Truelove, N., and Preziosi, R. (2016). A Galaxy-based bioinformatics pipeline for optimised, streamlined microsatellite development from Illumina next-generation sequencing data. *Conservation Genetics Resources*, 8(4):481–486.

Ley, S. (2019). Great barrier reef outlook report 2019. Australian Government: Great Barrier Reef Marine Park Authority. Accessed online: http://www.gbrmpa.gov.au/our-work/outlook-report-2019, 01/10/2019.

Li, Y., Korol, A., Fahima, T., and Nevo, E. (2004). Microsatellites within genes: structure, function and evolution. *Molecular Biology and Evolution*, 21(6):991–1007.

Liu, L., Dybvig, K., Panangala, V., van Santen, V., and French, C. (2000). GAA trinucleotide repeat region regulates M9/pMGA gene expression in *Mycoplasma gallisepticum*. *Infection and Immunity*, 68(2):871–876.

Nielsen, E., Cariani, A., Aoidh, E., Maes, G., Milano, I., Ogden, R., Taylor, M., Hemmer-Hansen, J., Babbucci, M., Bargelloni, L., Bekkevold, D., Diopere, E., Grenfell, L., Helyar, S., Limborg, M., Martinsohn, J., McEwing, R., Panitz, F., Patarnello, T., Tinti, F., Van Houdt, J., Volckaert, F., Waples, R., and Carvalho, G. (2012). Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. *Nature Communications*, 3.

Paris, J., Stevens, J., and Catchen, J. (2017). Lost in parameter space: a road map for STACKS. *Methods in Ecology and Evolution*, 8(10).

Pinar, G., Poyntner, C., Lopandic, K., Tafer, H., and Sterflinger, K. (2020). Rapid diagnosis of biological colonization in cultural artefacts using the MinION nanopore sequencing technology. *Internation Biodeterioration and Biodegradation*, 148(104908).

Ribout, C., Villers, A., Ruault, S., Bretagnolle, V., Picard, D., Monceau, K., and Gauffre, B. (2019). Fine-scale genetic structure in a high dispersal capacity raptor, the montagu's harrier (*Circus pygargus*), revealed by a set of novel microsatellite loci. *Genetica*, 147(1):69–78.

Semmouri, I., De Schamphelaere, K., Mees, J., Janssen, C., and Asselman, J. (2020). Evaluating the potential of direct RNA nanopore sequencing: Metatranscriptomics highlights posible seasonal differences in a marine pelagic crustacean zooplankton community. *Marine Environmental Research*, 104836.

Slate, J. and Pemberton, J. (2007). Admixture and patterns of linkage disequilibrium in a free-living vertebrate population. *Journal of Evolutionary Biology*, 20(4):1415–1427.

Sodhi, N. and Ehrlich, P. (2010). *Conservation Biology for All*. Oxford University Press, Great Clarendon Street, Oxford, UK.

Sunderland, T., Sunderland-Groves, J., Shanley, P., and Campbell, B. (2009). Bridging the gap: How can information access and exchange between conservation biologists and field practitioners be improved for better conservation outcomes? *Biotropica*, 41:549–554.

Tautz, D. and Renz, M. (1984). Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Research*, 12(10):4127–4138.

Treco, D. and Arnheim, N. (1986). The evolutionary conserved repetitive sequence $d(tg.ac)_n$ promoted reciprocal exchange adn generates unusual recombinant tetrads during yeast meiosis. *Molecular and Cellular Biology*, 6(11):3934–47.

Vautard, R., Oldenborgh, G., Haustein, K., Otto, F., Vogel, M., Senevirante, S., Soubeyroux, J., Schneider, M., Drouin, A., Ribes, A., Kreienkamp, F., and Aalst, M. (2019). Human contribution to the record-breaking July 2019 heatwave in western Europe. World Weather Attribution. Accessed online: https://www.worldweatherattribution.org/human-contribution-to-the-record-breaking-july-2019-heat-wave-in-western-europe/, 14/10/2019.

Venter, J., Adams, M., Sutton, G., Kerlavage, A., Smith, H., and Hunkapiller, M. (1998). Shotgun sequencing of the human genome. *Science*, 280(5369):1540–1542.

Vieira, M., Santini, L., Diniz, A., and Munhoz, C. (2016). Microsatellite markers: what they mean and why they are so useful. *Genetics and Molecular Biology*, 39(3):312–328.

Wang, G., Chai, X., Zhang, J., Yang, W., Jiang, C., Chen, K., and Xiong, J. (2020). A strategy for complete telomere-to-telomere assembly of ciliate macronuclear genome using ultra-high coverage Nanopore data. *bioRxiv*.

Appendices

7.1 Appendix 1 - Published version of Chapter 2.

Chapter 2 was published in a modified form in the Journal of Zoo and Aquarium Research. A copy of the printed article is below.

JZAR JOURNAL OF ZOO AND AQUARIUM RESEARCH
www.jzar.org

Research article

Bespoke markers for ex-situ conservation: application, analysis and challenges in the assessment of a population of endangered undulate rays

Graeme Fox[1,2], Iulia Darolti[3,2], Jean-Denis Hibbitt[4], Richard F. Preziosi[1,2], John L. Fitzpatrick[5,2], Jennifer K. Rowntree[1,2]

[1]School of Science and the Environment, Manchester Metropolitan University, John Dalton East, Manchester, M1 5GD, UK
[2]Faculty of Life Sciences, The University of Manchester, Manchester, M13 9PT, UK
[3]Department of Genetics, Evolution and Environment, University College London, London, WC1E 6BT, UK
[4]SEA LIFE Global, Merlin Animal Welfare and Development, Lodmoor Country Park, Weymouth SEA LIFE Adventure Park, Preston Road, Weymouth, Dorset, DT4 7SX, UK.
[5]Department of Zoology/Ethology, Stockholm University, Stockholm, SE-106 91, Sweden

Correspondence: Dr Jennifer K. Rowntree; j.rowntree@mmu.ac.uk

Keywords: elasmobranchii, microsatellite markers, next generation sequencing, population genetic structure, *Raja undulata.*

Article history:
Received: 14 Mar 2017
Accepted: 15 Mar 2018
Published online: 30 Apr 2018

Abstract
Genetic data are important and informative in the management of ex-situ populations. Where the risk of inbreeding is particularly great, it is critical that tools are employed that allow for the quantification of genetic variation and to identify potential breeding pairs. This study demonstrates the rapid application of laboratory and bioinformatics techniques to develop a novel microsatellite marker panel for use with a population of the endangered undulate ray (*Raja undulata*) and shows how a minimally invasive sampling method can be used with aquarium-dwelling individuals. The study assesses the population and investigates how informative a small microsatellite marker panel is to the conservation of a restricted ex-situ group. It was found that after a single captive generation of *R. undulata* there is no detectable evidence of reduced heterozygosity and no observable aquaria effects or differences between the generations. In conclusion, the study demonstrates that it is practical, quick and informative to develop a bespoke panel of markers to aid ex-situ conservation efforts of non-model species and make recommendations that these processes should constitute the minimum effort required in managing such a population.

Introduction

The elasmobranchii are a subclass of carnivorous, cartilaginous fish, including the sharks, rays, skates and sawfish. These species are found extensively in coastal, demersal and pelagic marine habitats and an additional minority inhabit freshwater systems (Compagno 1990). Common traits include slow growth and low productivity (Frisk et al. 2001; Walker 1998), resulting in high vulnerability and slow response to overexploitation from fishing activities (Ferretti et al. 2010; Smith et al. 1998). Recorded declines in elasmobranch populations over recent decades are typically associated with increasing fishing effort; an effect which can be seen in oceans the world over, for example in the Gulf of Mexico (Shepherd and Myers 2005); the Northwest Atlantic (Baum et al. 2003); the Mediterranean Sea (Ferretti et al. 2008); the Sea of Japan (Nakano 1999) and the Indian Ocean (Appukuttan and Nair 1988). Whether fishing effort targets elasmobranchs specifically (Rose 1998; Stevens et al. 2000) or they are a common feature of bycatch (Oliver et al. 2015), with the majority of global fisheries at risk of overexploitation (Botsford et al. 1997) the long-term effect on elasmobranch populations is largely unknown (Baum et al. 2003).

The undulate ray (*Raja undulata*) is an endangered skate often present in bycatch of commercial trawl fishing operations off the south coast of England, France, western Ireland and southern Portugal (Coehlo et al. 2009). Existing in small isolated populations, the species has recorded declines of up to 80% in some areas since the early 1980s, which has been directly attributed to fishing activities (Ellis et al. 2012). In 2009, the species was classified as endangered by the IUCN (Gibson et al. 2008). A managed breeding and monitoring programme

(Mon-P) was established in 2010 by the European Association of Zoos and Aquaria (EAZA) in response to the new IUCN classification and a European Union ban on the landing of this skate species was put in place. Currently, 36 aquaria across nine countries hold *R. undulata*. As part of the larger European breeding program, a small captive group is maintained across several UK aquaria, comprising a mixture of wild-caught and captive-bred individuals. Very little is known about the genetic diversity or population genetic structure of this species either in captivity or in the wild. The elasmobranchii are a charismatic focal point of interest for the general public in aquaria and are the subject of intense conservation effort to manage their ex-situ conservation. With >100 chondrichthyan species present in European zoos and aquaria (8.6% of all known elasmobranch species), there is great interest in the community for methods and techniques for sustainable conservation of these animals (Janse et al. 2017).

Non-random mating and genetic drift are major concerns for small populations and can have devastating implications for the evolutionary potential of the group. The small size of the population limits potential reproductive pairings, as inbreeding becomes a risk with the increased probability of a pair of individuals being related to one another (Witzenberger and Hochkirch 2011). Prolonged inbreeding in a closed population increases the probability of progeny being homozygous at a given locus, resulting in the overall reduction of heterozygosity of the group after successive generations. Genetic drift and adaptation to captivity can also contribute to the loss of rare alleles and overall reduction in heterozygosity (Price and Hadfield 2013; Willoughby et al. 2014). It is widely recognised that the fitness of a population is inversely related to allelic homozygosity, and severe effects, such as loss of viability or infertility, can present after just a few generations of close inbreeding (Frankham et al. 2004). These detrimental effects are cumulative as they are amplified by successive generations in captivity (Christie et al. 2012). As a result, the longer it has been in isolation, the less-well suited a captive population becomes to providing individuals for release (Earnhardt 2010; Lacy 2012). It is imperative, therefore, that the genetic variation present at the founding of the ex-situ population be carefully retained and inbreeding avoided through strategic genetic management of the population (Fernández et al. 2004; Frankham et al. 2010; Pelletier et al. 2009).

Under ideal conditions, during the establishment of a new ex-situ population, the entire group should be assessed using genetic markers to estimate the diversity of the cohort and help establish a baseline of genetic diversity, to identify any genetic similarity of founding individuals and to support future management. In the case of an existing population, genetic markers should be used even in the presence of detailed keeper reports and pedigrees; whilst these resources contain valuable information, they are limited in scope to the time that the individuals (or their ancestors) have been known to the relevant managers. The most common genetic marker used in analyses of this type is the microsatellite; short, repetitive, hypervariable regions of DNA that appear to be a feature universal to all genomes. Microsatellite marker panels are available in online databases for many species and published, optimised methodologies are available for developing novel sets of markers (Castoe et al. 2012; Griffiths et al. 2016). As the rate of species extinction is elevated above the background rate (Pimm et al. 2014) and there is potential for an unprecedented increase in the number of ex-situ populations being managed across a wide range of taxa (Dawson et al. 2011), it is imperative that general best practice guidelines in genetic management are established now. In line with the recommendations of Witzenberger and Hochkirch (2011) and Janse et al. (2017), the current best practice is argued to be the use genetic markers to characterise the diversity and relatedness of individuals in a captive breeding program and this should be the minimum standard required for the establishment, or maintenance, of any ex-situ conservation programme.

When sampling for the collection of DNA, the aim should be to minimise stress or discomfort experienced by the subject whilst collecting high-quality genomic template, especially in the case of an endangered or threatened species. Tissue sampling or destructive biopsy is clearly counterproductive in some cases, therefore the development and testing of non- or minimally invasive sampling methods is paramount. Here, a minimally invasive sampling method, developed for wild elasmobranchs by Lieber et al. (2013), is tested on aquarium specimens and found to be highly successful when combined with an off-the-shelf DNA extraction kit that enables isolation of high-purity DNA from the mucus layer.

In this investigation, bioinformatics techniques are used to develop a novel microsatellite marker panel suitable for use in *Raja undulata*, using Illumina shotgun next-generation sequencing data. These markers are then optimised in the laboratory and used to characterise a small ex-situ population. The viability and confidence with which the small marker panel can be used for population management is assessed, whilst providing a snapshot of the diversity contained within this population of captive elasmobranchs.

Methods

Microsatellite marker development

High-throughput, shotgun genomic sequencing can be used in order to identify microsatellite regions in the target genome. High quality, large molecular weight, genomic DNA is essential for successful next-generation sequencing and can be collected in a variety of ways, often using a species-specific method. Samples of blood, tissue or buccal swabs (Dunn et al. 2010) are also commonly used for genetic sampling. In this instance, tissue samples were obtained from a female ray that had been euthanised due to terminal ill health resulting from a severe fungal infection of the lateral line system. A range of tissue samples were taken from the animal post euthanasia under the guidance of Mark F. Stidworthy, veterinary pathologist at International Zoo Veterinary Group (IZVG). DNA was extracted from 25 mg heart tissue using the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany), following the manufacturer's protocol and checked for quality on a NanoDrop ND-1000 spectrophotometer (260/280 >1.4) and on a 1% agarose electrophoresis gel. A sequencing library was prepared using 50 ng genomic DNA and analysed on an Illumina MiSeq platform at the University of Manchester (UK) Genomics Facility using a shotgun, paired-end 2*250 sequencing methodology (Nextera DNA Library Preparation Kit, Illumina, San Diego, USA). In total, 11,019,590 raw sequencing reads were produced from the MiSeq run. Low quality regions were removed from each end of the reads, reads were trimmed using the average quality score over a sliding-window of 4 nt and a quality threshold of 20, and a minimum length of 50 nt was applied using Trimmomatic v0.0.4 (Bolger et al. 2014). If either of the paired-end reads failed a quality check, both reads were discarded, thus maintaining parity in the paired-end data. A majority (92%) of reads successfully passed quality filtering and were subsequently screened for potential microsatellite loci using pal_finder v0.02.04 software (Castoe et al. 2012). Non-perfect repeat loci were discarded and a minimum motif size of 3 nt was implemented (Griffiths et al. 2016).

Primer sequences were designed using Primer3 v.4.0.0 (Koressaar and Remm 2007; Untergrassar et al. 2012) using conditions optimised for the Qiagen Type-it microsatellite PCR kit (Qiagen, Hilden, Germany) (optimum length: 25 nt, minimum length: 18 nt, maximum length: 30 nt, minimum GC%: 45%, maximum GC%: 65%, minimum melting temperature: 62°C, maximum melting

temperature: 75°C, optimum melting temperature: 68°C, with remaining options set to the Primer3 default values); a set of PCR reagents designed specifically for amplification of microsatellite loci. The pal_finder process produced 698 potential loci that were ranked by predicted utility as a microsatellite marker (larger motifs preferred) and the primer sequences from the first 24 results were used to purchase DNA oligos from Sigma Aldrich (Missouri, USA) (scale: 0.025 µmole, purification: DST).

Sampling

For characterisation of the microsatellite loci, the 35 captive *R. undulata* (17 wild caught, 18 captive bred) were sampled using a modified form of the minimally invasive sampling method developed for wild elasmobranch sampling by Lieber and colleagues (2013), a method not known to have been previously demonstrated on captive animals. Small (1.5 cm x 2.5 cm), autoclaved sections of kitchen scouring pad (Vale Mill Ltd., Rochdale) were used to gently scrub the pectoral fin of the rays against the direction of the scales removing epidermal mucous secretions. Inter-species contamination was controlled, to the best of our ability, through the use of the species-specific PCR primers. As the markers were designed in a sample taken from excised heart tissue of an undulate ray (low risk of contamination), successful marker amplification implies a lack of contamination as the target DNA was of the same taxa as the heart sample. Intra-species contamination is more difficult to control for; however, it appears not to have been an issue, as microsatellite peak traces did not show multiple banding. The pads were immediately placed into individual tubes of absolute ethanol and stored at −80°C. During DNA extraction, extraneous pad was removed and DNA was extracted using the E.Z.N.A. Mollusc DNA Kit (Omega Bio-Tek, Norcross, USA); the use of chloroform:isoamyl alcohol (24:1) successfully isolating the mucus, precipitating proteins and producing high quality DNA extract. Elution was performed in 100 uL MilliQ water and used in downstream PCR for genotyping. This sampling technique reduces stress and damage to the animal as it minimises, or eliminates in some cases, the time the specimen spends out of the water during sampling. The technique could potentially be applicable to any captive elasmobranch with a mucus layer on the skin. A total of 35 animals were sampled from 10 different aquaria. More details as to the provenance of the samples are given in Table 1. Samples were also taken from several related *Raja* species (*R. microcellata, R. brachyura, R. montagui* and *R. clavata*) in order to test the cross-compatibility of the primers.

Marker amplification

Twenty-four potential markers were tested in the laboratory, of which eight successfully amplified.

PCR amplifications of 5 uL total volume were performed using the Qiagen Type-it Microsatellite PCR Kit (Qiagen, Hilden, Germany). Reactions consisted of 2.5 µL Type-it mastermix, 1.5 µL PCR grade H_2O, 0.5 µL genomic DNA at 20 ng/µL and 0.5 µL primer pair at 2 µM. This 5 µL reaction was amplified under the conditions specified by the PCR kit (5 min 95°C, 28x {30 sec 95°C, 90 sec 60°C, 30 sec 72°C}, 30 min 60°C) and successful amplifications were confirmed by the presence of bands on a 1% agarose electrophoresis gel. A three-primer universal-tailed approach was used to label amplicons with fluorescent moieties (Blackett et al. 2012) and fragment length reported using an Applied Biosystems 3730 DNA analyser capillary sequencer (Applied Biosystems, Foster City, California, USA) and GeneScan 500 LIZ dye size standard (Thermo Fisher Scientific, Carlsbad, USA) at the University of Manchester DNA Sequencing Facility.

Population genetic analysis

Raw data analysis was performed using GeneMapper 5.0 (Thermo Fisher Scientific, Carlsbad, USA) and confirmed that loci were scoreable and polymorphic. The novel markers were analysed for evidence of linkage disequilibrium and for Hardy–Weinberg

Table 1. Sample numbers taken from each of the 10 UK aquaria. The provenance of each sample is given as well as the number of private alleles detected at each aquarium.

Aquarium	Total N	Wild Caught	Captive Bred	Private Alleles
Sea Life London Aquarium	9	4	5	2
Weymouth Sea Life Adventure Park	5	1	4	2
Sea Life Blackpool	4	2	2	0
Sea Life Chessington	5	1	4	0
Sea Life Adventure, Southend	3	2	1	1
Sea Life Great Yarmouth	2	2	0	2
Sea Life Loch Lomond	1	1	0	0
Blue Reef Aquarium, Portsmouth	3	3	0	1
National Marine Aquarium, Plymouth	2	0	2	3
Marine Biology Association, Plymouth	1	1	0	0

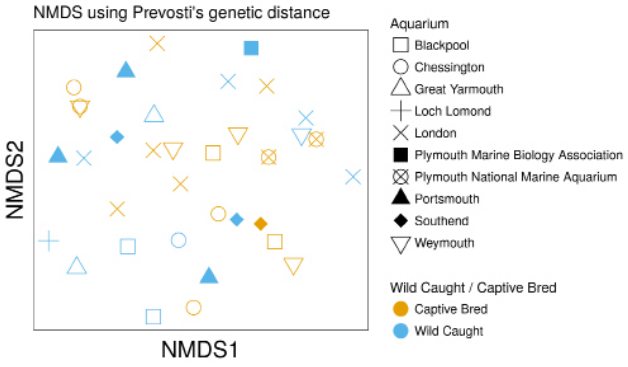


Figure 1. Ordination of Prevosti's genetic distance between individuals derived via non-metric multidimensional scaling (NMDS). Each point represents an individual, point shapes are aquaria and point colour represents whether the individual is wild caught or captive bred. The stress value of the NMDS was 0.17, demonstrating reasonable confidence in the ordination whilst maintaining a minimum number of dimensions.

Table 2. Locus ID, nucleotide motif, number of alleles (NA), size range of fragments (SR), PCR annealing temperature (TA), expected (Hexp) and observed (Hobs) heterozygosity, number of individuals tested (N), P-value from testing for Hardy–Weinberg equilibrium (PHWE) and primer nucleotide sequences (5' to 3' orientation). *Marker RU13 not used in this study due to deviation from expected HWE values.

Locus Name	Motif	NA	SR (bp)	TA (°C)	Hexp	Hobs	PHWE	N	Primer Sequences (5' - > 3' orientation)	GenBank Accession
Ru02	AAGAGG	10	347 - 419	60	0.808	0.800	0.0180	35	CCCTGTTCTCCTGCTCTCCATTACC CTCTCCCTATAGCTCAGGCCTTCGG	MH049873
Ru03	ACTGCC	10	412 - 463	60	0.827	0.882	0.0694	34	CATTCACAACTGCAGTCCAATGTCC TCTGCTGTCAAGCTGTTGTGTCAGG	SRP134840
Ru08	AGGTG	13	351 - 415	60	0.887	0.800	0.0113	35	TGAGGAATTCATTGCCACAAACTGC TCCTCTCACATAACCCTGTGTATGCC	MH049874
Ru09	ATAG	22	209 - 385	60	0.945	0.939	0.1463	33	TCTTTGCTCCTACCGGTTCTTCTCG CAGAACAAGGCTTGGTGGTCTTGG	MH049875
Ru13*	ACAG	9	317 - 373	60	0.787	0.313	0	32	CATTCTTAACAGGGCAGCTACTTGTGG AAAGATTGGTAGGAAGATGGATCGG	MH049876
Ru14	AGGC	8	277 - 313	60	0.754	0.882	0.7937	34	ACCTCGAAACCGCCATTAAGAATCC CTGCATGTTATCGAGCAATCAGTCG	MH049877
Ru20	ACAG	9	374 - 407	60	0.846	0.886	0.1317	35	GGACACTTGACACAGCTTTGGTCTCC GGGAGTTACCTTCATGGTGAGACAGG	MH049878
Ru21	AAT	5	373 - 388	60	0.682	0.543	0.1631	35	CATGACTGGGGCTAGAAGGTGTTGC GTTAGAGCAGTCCGCCATGAAGGG	MH049879

equilibrium using GenePop v.4.2 online (Raymond and Rousset 1995; Rousset 2008). Estimates of pairwise relatedness were calculated for every pair of individuals using the triadic likelihood estimator of relatedness, a measure suited to a relatively small number of markers, implemented in Coancestry using the R (R Core Team, 2017) package "related" (Pew et al. 2015). The rate of heterozygosity, inbreeding coefficient and measures of genetic distance were calculated using the "adegenet" package in R (Jombart 2008; Jombart and Ahmed 2011; Rogers 1972). Rates of allelic richness and private alleles were identified using the R package "PopGenReport" (Adamack and Gruber 2014). The data were split by generation, and comparisons were drawn between each generation. In this instance, all wild-caught individuals were compared to all captive-bred offspring, as at the time of sampling there was only a single generation captive population (F1 generation).

Results

Eight polymorphic microsatellite markers were initially characterised and every marker demonstrated to amplify consistently at an annealing temperature of 60°C, advantageous for multiplex PCR. These novel markers were used to genotype 35 captive *R. undulata* individuals at the eight loci. GENEPOP results for linkage disequilibrium (LD) showed that 48% of total marker pairs exhibited significant evidence of LD; however, when just the wild-caught individuals were tested, this percentage was reduced to zero. GENEPOP was also used to check for deviation from the expected allele frequencies of Hardy–Weinberg. Three markers showed significant deviation in the total population and a single marker (Ru13) showed deviation from expected frequencies in the wild-caught animals only. This marker (Ru13) was subsequently removed from the analysis. Summary statistics for the remaining seven markers are given below in Table 2. A success rate of 98% was achieved in obtaining genotypic data. Average allelic richness was 7.0 in the wild-caught group, 6.4 in the captive-bred group and 1.7 per aquarium. The average observed rate of heterozygosity at each marker was 0.81. Observed heterozygosity (Hobs) and the average estimated inbreeding coefficient (r) were calculated for the wild-caught animals (Hobs=0.80, r=0.21±0.003) and the first generation, captive-bred individuals (Hobs=0.83, r=0.18±0.005).

Table 3. Microsatellite markers tested in several other *Raja* species. Size ranges in a limited number of samples.

Species	Locus Name							
	Ru02	Ru03	Ru08	Ru09	Ru13	Ru14	Ru20	Ru21
Raja microcellata								
	341-419	412-463	351-432	209-385	317-373	277-376	374-407	373-388
Raja brachyura								
	347-377	408-463	351-428	204-385	317-419	277-391	374-407	373-388
Raja montagui								
	347-364	412-483	351-415	209-385	317-373	277-313	374-422	373-388
Raja clavata								
	343-419	412-463	351-415	209-385	285-373	277-313	374-407	373-388

There was no significant difference in either heterozygosity (two sample t-test, t=0.52644, df=10.171, P=0.6099) or the average inbreeding coefficient (two sample t-test: t=−1.0356, df=14.225, P=0.3177) between wild-caught and captive-bred individuals. One to three private alleles were discovered in six of the 10 aquaria (aquarium population size ranging from 1–9 individuals). A nonmetric multidimensional scaling (NMDS) analysis of Provesti's genetic distance among individuals (Figure 1), calculated using the R (R Core Team, 2018) package "vegan" (Oksanen et al. 2017), provides a visual interpretation of the genetic similarity of individuals. The calculated stress value of the NMDS was 0.17, the lowest stress value of each of the measures of genetic distance calculated using the "adegenet" (Jombart 2008; Jombart and Ahmed 2011) package in R. A stress value of <0.2 indicates a fair fit of the data in the NMDS analysis (Kruskal 1964).

The minimally invasive extraction method and the seven primer pairs were tested with several other species of the *Raja* genus (species listed previously) and were demonstrated to successfully amplify polymorphic loci in every species tested, suggesting good cross-species compatibility of the primers and sampling technique. Allelic range in these species very closely matched those discovered in *R. undulata* (Table 3). Four or fewer samples from each species were tested and, therefore, more detailed locus statistics are not provided here.

Discussion

The goal of this study was to develop and optimise a novel set of microsatellite markers for the endangered undulate ray (*Raja undulata*) and subsequently assess their power and informativeness for ex-situ conservation of this species. Genomic DNA, extracted from a tissue sample, was successfully used to generate a sequencing library, and bioinformatics and laboratory techniques were employed to discover and optimise seven microsatellite markers from the resulting next-generation sequencing (NGS) dataset. In order to undertake genetic analyses of this nature, a reliable source of DNA is required, but often this can come at the cost of distress or harm to the subject. Therefore, non-invasive genetic sampling methods are preferable to invasive tissue, blood or biopsy sampling, particularly for threatened species. Although an initial tissue sample was used for the development of the markers, a minimally-invasive sampling method for the collection of the remaining samples from the captive animals (Lieber et al. 2013) was tested. This technique takes advantage of the mucus secreted by the skin of many elasmobranchs and this study demonstrates the successful isolation of high-quality, amplifiable DNA from captive animals. The new markers were used to genotype a small captive population of 35 animals, across 10 UK aquaria, demonstrating that the minimally-invasive sampling methodology was suitable for a study of this nature. Several quality-checking procedures were applied to the markers themselves, such as tests for linkage disequilibrium (LD) or deviations from Hardy–Weinberg Equilibrium (HWE). Evidence of both LD and deviation from HWE was observed in some markers. The deviation from expected HWE can be attributed to the fact that the test population breaks many of the underlying assumptions of HWE, mainly that one should consider a large, unrelated population, which is not the case here. Several statistical analyses of the data were performed, making routine measurements of heterozygosity of the population at these loci, calculating inbreeding coefficients and genetic distance, for example.

The results show rates of heterozygosity at each marker ranging from 0.54–0.94 (average 0.81), implying that when all markers are taken into account, the rate of genetic variation in the captive population is not likely to be significantly lower than the wild population from which it was founded. For comparison, Chapman et al. (2011) used seven microsatellite markers to measure heterozygosity in an elasmobranch population consisting of 104 individuals of the critically-endangered smalltooth sawfish (*Pristis pectinata*) and discovered an average rate of heterozygosity of 0.83. Heterozygosity rates in wild-caught animals and captive-bred, F1 generation individuals did not show any significant difference, demonstrating that a high proportion of genetic variation has been carried into this generation. Data reporting the proportion of wild-caught individuals that successfully contributed to the F1 generation are unfortunately not available. These measures should be repeated at each new generation and can be interpreted as a proxy for the measure of total variation in the group. The captive-bred *R. undulata* of the present study had an average rate of heterozygosity of 0.83. It is important to note, however, that these results on the captive-bred population only take into account the F1 generation and that any decrease in the rate of heterozygosity will likely become apparent over subsequent generations (Willoughby et al. 2017). Continued monitoring via the methods explained in this study will be critical to continue to evaluate the genetic diversity of the population and to continue to monitor for inbreeding depression. Several aquaria housing private alleles within their cohort have been identified, and this information may be useful for maintaining genetic variation when the breeding plan is developed.

While it is common to calculate the likely pedigree (i.e. relatedness) from this type of genetic data, the power to correctly assign offspring to parents will be very low for captive populations with a limited captive population size. In these cases, it is far more informative to directly examine the genetic similarity of individuals. The calculation of Provesti's genetic distance (Prevosti et al. 1975) enabled the visualisation of a proxy measure of dissimilarity between individuals (see Figure 1) through calculating the absolute genetic distance between each pair of individuals. Figure 1 shows no clustering around a particular aquarium or between the wild caught or captive bred groupings, indicating the lack of an aquarium effect or differentiation of the F1 generation from the wild individuals. Rather, the individual genotypes suggest a homogenous mixture with no apparent groupings, or sub-structuring emerging. These results fall within expectations as ~50% of the total individuals were wild caught (17 of 35) and so can be expected to be reasonably unrelated to one another as they originate from a wild population. Progeny from relatively high admixture would be expected to maintain high levels of variation in the F1 generation and similarly be relatively unrelated to one another (with the exception of siblings, parents-progeny, etc.).

This study leads to the recommendation that similar analyses be performed as new individuals are caught, born or moved between aquaria to enable population managers to intervene should a particular group of individuals appear to become distinct from other groups, or when one of the measures, or proxy measures, of variation among individuals begins to fall. With a greater number of microsatellite markers, the work could be extended to include relatedness estimates of a much higher confidence and this would also lead to the production of accurate pedigrees—very useful tools to the community managing these animals, but beyond the scope of this piece of work.

Conclusion

Ex-situ conservation is a very important management tool and is likely to be increasingly used as the rate of anthropogenic-influenced species declines continues to climb (Ceballos et al. 2015). Captive populations must be carefully and strategically managed in order to successfully provide individuals for reintroduction, maintain genetic variation and reduce the negative effects of inbreeding (Frankham et al. 2004). Janse et al.

(2017) succinctly summarised the contemporary elasmobranch populations in European aquaria and identified the requirement for good programme management. This study demonstrates that researchers can move relatively quickly from collecting tissue/swab samples, through designing a novel marker panel to producing quantifiable, genetic data and drawing conclusions regarding the structure of a captive population (the majority of the work on this analysis was performed in a matter of a few months). In the absence of a good quality pedigree or studbook, these techniques should form the minimum requirement when working with ex-situ populations, and as NGS technologies continue to improve, the number and nature of available markers will also increase, leading to significant gains in the quality of the data available. The power of this particular study was limited by a lack of markers, thus preventing some analyses from being performed. However, from the data generated here, it is evident that the population of undulate rays in UK aquaria do not currently appear to be suffering from any malady resulting from their small population size, and the findings appear to fall in line with other managed groups of elasmobranchs. The results, however, constitute a time-bound observation and are therefore only representative of the population at the time the samples were taken. In conclusion, the study has shown that it is feasible and useful to design and optimise a panel of markers for a small, ex-situ population and that even with a small number of markers, the resulting data can be informative and help with the management of the population. With these markers available to the community, it is hoped that a better understanding of the captive population in UK aquaria in relation to individuals in European aquaria and in wild populations can be reached. This study forms the basis for further scope of greater scope, encompassing a greater sample size, more sampling sites (aquaria) and more microsatellite markers to increase the statistical power of the analyses.

Acknowledgements

This research was funded by the University of Manchester Faculty of Life Sciences (FLS), a FLS Business Development Small Award, and the Sea Life Trust. Our thanks go to the DNA Sequencing Facility and the Genomic Technologies Core Facility, both at the University of Manchester (UK), for their expert advice and services and two anonymous reviewers for their helpful comments, which improved the manuscript.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Adamack A.T., Gruber B. (2014) PopGenReport: simplifying basic population genetic analyses in R. *Methods in Ecology and Evolution* 5: 384–387.

Appukuttan K.K., Nair K.P. (1988) Shark resources of India, with notes on biology of a few species. In: Joseph, M.M. (ed.). *The First Indian Fisheries Forum, Proceedings*. Mangalore, India: Asian Fisheries Society, Indian Branch, 173–184.

Baum J.K., Myers R.A., Kehler D.G., Worm B., Harley S.J., Doherty P.A. (2003) Collapse and conservation of shark populations in the Northwest Atlantic. *Science* 299: 389–392.

Blackett M.J., Robin C., Good R.T., Lee S.F., Miller A.D. (2012) Universal primers for fluorescent labelling of PCR fragments—an efficient and cost effective approach to genotyping by fluorescence. *Molecular Ecology Resources* 12: 456–463. doi: 10.1111/j.1755-0998.2011.03104.x.

Bolger A.M., Lohse M., Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120. doi: 10.1093/bioinformatics/btu170.

Botsford L.W., Castilla J.C., Peterson C.H. (1997) The management of fisheries and marine ecosystems. *Science* 277: 509–515.

Castoe T.A., Poole A.W., de Koning A.P.J., Jones K.L., Tomback D.F., Oyler-McCance S.J., Fike J.A., Lance S.L., Streicher J.W., Smith E.N., Pollock D.D. (2015) Correction: Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PLOS ONE* 10: e0136465. doi: 10.1371/journal.pone.0136465.

Ceballos G., Ehrlich P.R., Barnosky A.D., García A., Pringle R.M., Palmer T.M. (2015) Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Science Advances* 1. doi: 10.1126/sciadv.1400253.

Chapman D.D., Simpfendorfer C.A., Wiley T.R., Poulakis G.R., Curtis C., Tringali M., Carlson J.K., Feldheim K.A. (2011) Genetic diversity despite population collapse in a critically endangered marine fish: The smalltooth sawfish (Pristis pectinata). *Journal of Heredity* 102: 643–652. doi:10.1093/jhered/esr098.

Christie M.R., Marine M.L., French R.A., Blouin M.S. (2012) Genetic adaptation to captivity can occur in a single generation. *Proceedings of the National Academy of Sciences* 109: 238–242.

Coelho R., Bertozzi M., Ungaro N., Ellis J. (2009) *Raja undulata. The IUCN Red List of Threatened Species* 2009: e.T161425A5420694. http://dx.doi.org/10.2305/IUCN.UK.2009-2.RLTS.T161425A5420694.en. Downloaded on 27 October 2017.

Compagno L.J.V. (1990) Alternative life-history of cartilaginous fishes in time and space. *Environmental Biology of Fishes* 28: 33–75.

Dawson T.P., Jackson S.T., House J.I., Prentice I.C., Mace G.M. (2011) Beyond Predictions: Biodiversity Conservation in a Changing Climate. *Science* 332(6025): 53–58. DOI:10.1126/science.1200303.

Dunn S.J., Barnowe-Meyer K.K., Gebhardt K.J., Balkenhol N., Waits L.P., Byers J.A. (2010) Ten polymorphic microsatellite markers for pronghorn (Antilocapra americana). *Conservation Genetics Resources* 2: 81–84.

Earnhardt J.M. (2010) The role of captive populations in reintroduction programs. In: Kleiman D.G., Thompson K.V., Baer C.K. (eds.). *Wild mammals in captivity: principles and techniques for zoo management.* 2nd edition. Chicago, IL, University of Chicago Press, 268–280.

Ellis J.R., McCully S.R., Brown M.J. (2012) An overview of the biology and status of undulate ray *Raja undulata* in the North-east Atlantic Ocean. *Journal of Fish Biology* 80: 1057–1074.

Estes J.A., Burdin A., Doak D.F. (2015) Sea otters, kelp forests, and the extinction of Steller's sea cow. *Proceedings of the National Academy of Sciences* 113: 880–885.

Fernández J., Toro M.A., Caballero A. (2004) Managing individuals' contributions to maximise allelic diversity maintained in small, conserved populations. *Conservation Biology* 18: 1358–1367.

Ferretti F., Myers R.A., Serena F., Lotze H.K. (2008) Loss of large predatory sharks from the Mediterranean sea. *Conservation Biology* 22: 952–964.

Ferretti F., Worm B., Britten G.L., Heithaus M.R., Lotze H.K. (2010) Patterns and ecosystem consequences of shark declines in the oceans. *Ecology Letters* 13: 1055–1071. doi: 10.1111/j.1461-0248.2010.01489.x.

Frankham R., Ballou J.D., Briscoe D.A. (2004) *A Primer of Conservation Genetics*. Cambridge, UK, Cambridge University Press.

Frankham R., Ballou J.D., Briscoe D.A. (2010) *Introduction to conservation genetics,* 2nd edn. Cambridge, UK, Cambridge University Press.

Frisk M.G., Miller T.J., Fogarty M.J. (2001) Estimation and analysis of biological parameters in elasmobranch fishes: a comparative life history study. *Canadian Journal of Fisheries and Aquatic Sciences* 58: 969–981.

Gibson C., Valenti S.V., Fowler S.L., Fordham S.V. (2006) *The Conservation Status of Northeast Atlantic Chondrichthyans; Report of the IUCN Shark Specialist Group Northeast Atlantic Regional Red List Workshop.* VIII + 76pp. IUCN SSC Shark Specialist Group.

Griffiths S.M., Fox G., Briggs P.J., Donaldson I.J., Hood S., Richardson P., Leaver G.W., Truelove N.K., Preziosi R.F. (2016) A Galaxy-based bioinformatics pipeline for optimised, streamlined microsatellite development from Illumina next-generation sequencing data. *Conservation Genetic Resources* 8: 481–486.

Ivy J.A., Putnam A.S., Navarro A.T., Gurr J., Ryder O.A. (2016) Applying SNP-derived molecular co-ancestry estimates to captive breeding programs. *Journal of Heredity* 107(5): 403–12.

Janse M., Zimmerman B., Geerlings L., Brown C., Nagelkerke L.A.J. (2017) Sustainable species management of the elasmobranch populations within European aquariums: a conservation challenge. *Journal of Zoo and Aquarium Research* 5(1): 172–181.

Jombart T. (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24: 1403–1405. doi: 10.1093/bioinformatics/btn129.

Jombart T., Ahmed I. (2011) adegenet 1.3-1: new tools for the analysis of genome wide SNP data. *Bioinformatics* 27: 3070–3071.

Koressaar T., Remm M. (2007) Enhancements and modifications of primer design program Primer3. *Bioinformatics* 23: 1289–1291.

Kruskal J.B. (1964) Multidimensional scaling by optimising goodness of fit to a nonmetric hypothesis. *Psychometrika* 29(1).

Lacy R.C. (2012) Achieving true sustainability of zoo populations. *Zoo Biology* 32: 19–26. doi: 10.1002/zoo.21029.

Lieber L., Berrow S., Johnston E., Hall G., Hall J., Gubili G., Sims D.W., Jones C.S., Noble L.R. (2013) Mucus: aiding elasmobranch conservation through non-invasive genetic sampling. *Endangered Species Research* 21: 215–222.

Oksanen J., Blanchet F., Friendly M., Kindt R., Legendre P., McGlinn D., Minchin P.R., O'Hara R.B., Simpson G.L., Solymos P., Henry M., Stevens H.H., Szoecs E., Wagner H. (2017) *vegan: Community Ecology Package.* R package version 2.4-4 https://CRAN.R-project/package=vegan.

Oliver S., Braccini M., Newman S.J., Harvey E.S. (2015) Global patterns in the bycatch of sharks and rays. *Marine Policy* 54: 86–97.

Pelletier F., Reale D., Watters J., Boakes E.H., Garant D. (2009) Value of captive populations for quantitative genetics research. *Trends in Ecology and Evolution* 24: 263–270.

Pew J., Muir P.H., Wang J., Frasier T.R. (2015) related: an R package for analysing pairwise relatedness from codominant molecular markers. *Molecular Ecology Resources* 15: 557–561.

Pimm S.L., Jenkins C.N., Abell R., Brooks T.M., Gittleman J.L., Joppa L.N., Raven P.H., Roberts C.M., Sexton J.O. (2014) The biodiversity of species and their rates of extinction, distribution and protection. *Science* 344(6187). DOI: 10.1126/science.1246752.

Prevosti A., Ocaña , Alonso G. (1975) Distances between populations of *Drosophila subobscura*, based on chromosome arrangement frequencies. *Theoretical and Applied Genetics* 45(6): 231–241.

Price M.R., Hadfield M.G. (2013) Population genetics and the effects of a severe bottleneck in an ex situ population of critically endangered Hawaiian tree snails. *PLOS ONE* 9: e114377. doi:10.1371/journal.pone.0114377.

R Core Team (2018) *R: A language and environment for statistical computing.* F Foundation for Statistical Computing, Vienna, Austria. URL https://R-project.org/.

Raymond M., Rousset F. (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity* 86: 248–249.

Rogers J.S. (1972) Measures of genetic similarity and genetic distance. In: *Studies in Genetics* VII, Austin, Texas, University of Texas Publication 7213, 145–153.

Rose D.A. (1998) *Shark fisheries and trade in the Americas*, volume 1: North America. Washington D.C., USA: TRAFFIC.

Rousset F. (2008) Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. *Molecular Ecology Resources* 8: 103–106.

Shepherd T.D., Myers R.A. (2005) Direct and indirect fishery effects on small coastal elasmobranchs in the northern Gulf of Mexico. *Ecology Letters* 8: 1095–1104. doi: 10.1111/j.1461-0248.2005.00807.x.

Nakano H. (1999) Fishery management of sharks in Japan. In: Shotton, R. (ed). *Case studies of the management of elasmobranch fisheries.* Fisheries and Aquaculture Department, Food and Agriculture Organization of the United Nations, ISBN: 92-5-104291-8.

Smith S.E., Au D.W., Show C. (1998) Intrinsic rebound potentials of 26 species of Pacific sharks. *Marine and Freshwater Research* 49: 663–678.

Stevens J.D., Bonfil R., Dulvy N.K., Walker P.A. (2000) The effects of fishing on sharks, rays and chimaeras (chondrichthyans), and the implications for marine ecosystems. *ICES Journal of Marine Science* 57: 476–494.

Untergrasser A., Cutcutache I., Koressaar T., Ye J., Faircloth B.C., Remm M., Rozen S.G. (2012) Primer3 - new capabilities and interfaces. *Nucleic Acids Research* 40(15): e115. doi:10.1093/nar/gks596.

Walker P.A., Hislop J.R.G. (1998) Sensitive skates or resilient rays? Spatial and temporal shifts in ray species composition in the central and northwestern North Sea between 1930 and the present day. *ICES Journal of Marine Science* 55: 392–402.

Wang J. (2011) COANCESTRY: A program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Molecular Ecology Resources* 11: 141–145.

Weir B.S., Cockerham C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38(6): 1358–1370.

Willoughby J.R., Fernandez N.B., Lamb M.C., Ivy J.A., Lacy R.C., DeWoody J.A. (2014) The impacts of inbreeding, drift and selection on genetic diversity in captive breeding populations. *Molecular Ecology* 11: 98–110. doi: 10.1111/mec.13020.

Willoughby J.R., Ivy J.A., Lacy R.C., Doyle J.M., DeWoody J.A. (2017) Inbreeding and selection shape genomic diversity in captive populations: Implications for the conservation of endangered species. *PLOS ONE* 12(4): e0175996.

Witzenberger K.A., Hochkirch A. (2011) Ex situ conservation genetics: a review of molecular studies on the genetic consequences of captive breeding programmes for endangered animal species. *Biodiversity and Conservation* 20: 1843–1861. doi: 10.1007/s10531-011-0074-4.

7.2 Appendix 2 - Published version of Chapter 3.

Chapter 3 was published in a modified form in the Molecular Ecology Resources journal. A copy of the printed article is below.

Received: 6 July 2018 | Revised: 17 June 2019 | Accepted: 18 June 2019

DOI: 10.1111/1755-0998.13065

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Multi-individual microsatellite identification: A multiple genome approach to microsatellite design (MiMi)

Graeme Fox[1] | Richard F. Preziosi[1] | Rachael E. Antwis[2] | Milena Benavides-Serrato[1,3] | Fraser J. Combe[1,4] | W. Edwin Harris[1,5] | Ian R. Hartley[6] | Andrew C. Kitchener[7] | Selvino R. de Kort[1] | Anne-Isola Nekaris[8] | Jennifer K. Rowntree[1]

[1]Ecology and Environment Research Centre, Department of Natural Sciences, Manchester Metropolitan University, Manchester, UK

[2]School of Environment and Life Sciences, University of Salford, Salford, UK

[3]Universidad Nacional de Colombia, Playa Salguero, Colombia

[4]Division of Biology, Kansas State University, Manhattan, KS, USA

[5]Crop and Environment Sciences, Harper Adams University, Newport, UK

[6]Lancaster Environment Centre, Lancaster University, Lancaster, UK

[7]Department of Natural Sciences, National Museums Scotland, Edinburgh, UK

[8]Department of Social Sciences, Faculty of Humanities and Social Sciences, Oxford Brookes University, Oxford, UK

Correspondence
Jennifer K. Rowntree, Ecology and Environment Research Centre, Department of Natural Sciences, Manchester Metropolitan University, Manchester, UK.
Email: j.rowntree@mmu.ac.uk

Abstract

Bespoke microsatellite marker panels are increasingly affordable and tractable to researchers and conservationists. The rate of microsatellite discovery is very high within a shotgun genomic data set, but extensive laboratory testing of markers is required for confirmation of amplification and polymorphism. By incorporating shotgun next-generation sequencing data sets from multiple individuals of the same species, we have developed a new method for the optimal design of microsatellite markers. This new tool allows us to increase the rate at which suitable candidate markers are selected by 58% in direct comparisons and facilitate an estimated 16% reduction in costs associated with producing a novel microsatellite panel. Our method enables the visualisation of each microsatellite locus in a multiple sequence alignment allowing several important quality checks to be made. Polymorphic loci can be identified and prioritised. Loci containing fragment-length-altering mutations in the flanking regions, which may invalidate assumptions regarding the model of evolution underlying variation at the microsatellite, can be avoided. Priming regions containing point mutations can be detected and avoided, helping to reduce sample-site-marker specificity arising from genetic isolation, and the likelihood of null alleles occurring. We demonstrate the utility of this new approach in two species: an echinoderm and a bird. Our method makes a valuable contribution towards minimising genotyping errors and reducing costs associated with developing a novel marker panel. The Python script to perform our method of multi-individual microsatellite identification (MiMi) is freely available from GitHub (https://github.com/graemefox/mimi).

KEYWORDS

cost-effective marker development, high-throughput sequencing, in silico quality control, microsatellite design, polymorphic loci detection, short tandem repeat (STR)

1 | INTRODUCTION

Microsatellites, short tandem repeats (STRs) or short simple repeats (SSRs), are exceptionally polymorphic repetitive regions of DNA found throughout the genomes of both eukaryotic and prokaryotic species (Bhargava & Fuentes, 2010; Rose & Falush, 1998). High rates of polymorphism, along with codominance and Mendelian inheritance, make them ideal markers for use in studies of population genetics (Abdul-Muneer, 2014; Goldstein & Pollock, 1997). Microsatellites have been the most popular choice of genetic marker for several decades in ecology, conservation and evolutionary research, and are extensively used in contemporary studies of population genetics, parentage and kinship identification, evolutionary processes and genetic mapping (Ribout et al., 2019; Vieira, Santini, Diniz, & de Munhoz, 2016). Although single nucleotide polymorphism (SNP) markers have become increasingly popular markers for population genetics, microsatellites remain a common choice due to well-documented methodologies, ease of application, low equipment demands and well-developed statistical analyses. Furthermore, there remain scenarios where SNPs are not practical for use, or microsatellites are preferred (Zhan et al., 2016). For example, the management of captive populations has benefited enormously by the inclusion of genetic information (Fox et al., 2018; Witzenberger & Hochkirch, 2011), which must be continually updated as small numbers of new individuals are added to collections or produced through mating. In these cases, it is impractical to perform repeated SNP analyses on small numbers of samples due to the expense associated with next-generation sequencing (NGS) to acquire high coverage SNPs. Conversely, once a microsatellite panel has been developed, additional individuals can be genotyped using the existing markers very quickly, and at very low cost (Puckett, 2016). Where non-invasive sampling methods are required, for example because a species is of conservation concern (e.g., Fox et al., 2018), it may prove to be impossible to acquire sufficient high molecular weight DNA to perform NGS for SNP genotyping. In contrast, microsatellite analysis is forgiving of low DNA template input, and many contaminants that may disrupt NGS library preparation can simply be diluted out prior to amplification. A simple literature search in Google Scholar indicated the publication of approximately 2,000 new microsatellite marker panels in 2018, suggesting that microsatellites are still very popular genetic markers, and we predict they will continue to be used extensively in conservation and ecology well into the future.

Ecological and conservation studies are often focused upon non-model species for which genetic markers are not available. The combination of affordable NGS and freely available bioinformatics tools can be used to identify tens of thousands of potential markers in a matter of days. Where probes were once used to target repeat regions of genetic code (Bloor, Barker, Watts, Noyes, & Kemp, 2001), shotgun genome sequencing does not require any prior knowledge of the genome, and is considered a nontargeted approach (Davey et al., 2011). Instead, random fragments of genomic DNA are sequenced, a fraction of which include SSRs within the length of the sequencing read. Free, open source software packages are available to detect SSRs and design suitable PCR primers to amplify the appropriate region of the genome; often referred to as the "seq-to-SSR" approach (Castoe et al., 2015; Griffiths et al., 2016). These developments, and the increasing availability of NGS technology globally, brings microsatellite marker discovery within the reach of ever more research laboratories as the cost-per-base of NGS continues to decrease (Koboldt, Steinberg, Larson, Wilson, & Mardis, 2013; McPherson, 2014), even for applied, species-focused conservation research with limited funding. Thus, the development of bespoke microsatellite marker panels has become commonplace.

The use of microsatellite markers is reliant upon variation in PCR product fragment length, and therefore microsatellites must be amplifiable by PCR, and must contain fragment length altering polymorphisms within the repetitive stretch of SSR sequence. Despite improvements delivered by NGS, the optimisation of a bespoke microsatellite panel remains a time consuming and costly process, largely because the primer pair for each potential marker still requires manual laboratory confirmation of both successful amplification and the presence of multiple alleles at each locus (Bloor et al., 2001). Typically, the development of a microsatellite marker is performed through the discovery of a microsatellite locus in a single individual, followed by analysis of the locus in several more individuals to test for consistent amplification and variation in PCR fragment size (Abdelkrim, Robertson, Stanton, & Gemmell, 2009). The main contributors to the cost of developing a panel of microsatellite markers are the NGS reagents, PCR reagents, PCR oligos, capillary electrophoresis, size standards and staff time. Improvements that enable reductions in cost or time associated with marker development will contribute to microsatellite markers becoming more widely available to ecological and conservation researchers.

Here we present a new conceptual approach to microsatellite marker design, demonstrated with a new bioinformatics technique applied to seq-to-SSR workflows. This technique is designed to improve the rate at which loci that are identified can be successfully amplified by PCR and produce informative genotype data. The innovation in our approach is the incorporation of information from the genomes of multiple individuals. This allows the in silico detection of polymorphic loci and the detection of several other important characteristics of a putative microsatellite marker, which are only detectable through multiple genome analysis. We demonstrate that this method reduces the number of markers that must be tested for polymorphism in the laboratory, and achieves an improved rate of successful marker development. Furthermore, our methods also minimise factors known to increase allelic dropout and invalidate genotyping results based upon molecular weight of PCR fragments. We refer to this technique as multi-individual microsatellite identification (MiMi). Here, we develop microsatellite markers using MiMi in two species: the green sea urchin (*Psammechinus miliaris*) and the Eurasian blue tit (*Cyanistes caeruleus*). For comparison, we also present the success rates of microsatellite development in *P. miliaris* and

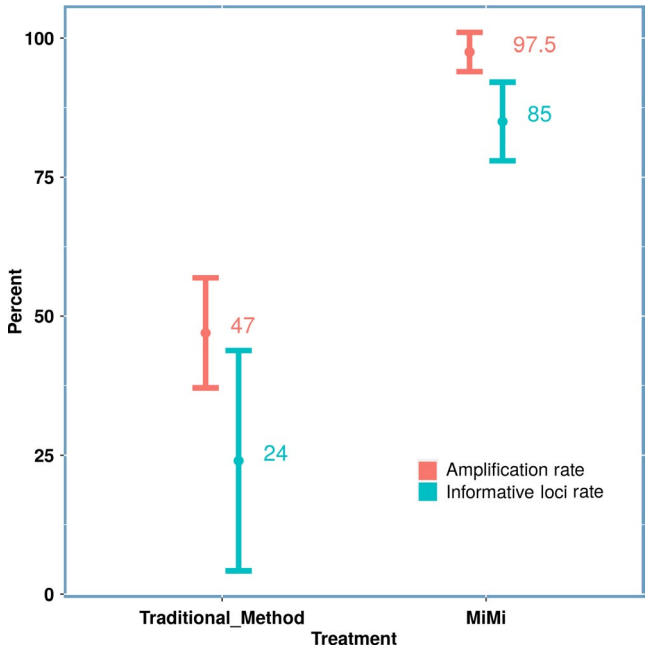


FIGURE 1 Summary statistics showing the rate at which potential microsatellite markers were successfully amplified in the laboratory, and the rate at which they were discovered to be informative. Markers were designed using both methodologies in *P. miliaris* and *C. caeruleus*. Stated values are the average for each design method, in each measure of success (amplification rate and informative loci rate). Error bars show the standard deviations. The use of MiMi results in both an increase in the rate at which markers amplify and are informative, and also a reduction in the variability at each of these measures compared to the traditional workflow

C. caeruleus, and in two other species (*Tragelaphus eurycerus isaaci* and *Nycticebus pygmaeus*), which were designed using a traditional microsatellite design method (Castoe et al., 2015; Griffiths et al., 2016). The results from the successful development of each panel of markers, combined with our refined bioinformatics method, provide a strong case for the utility of the MiMi concept and the value to microsatellite marker development.

2 | MATERIALS AND METHODS

2.1 | DNA extraction and sequencing

Prior to DNA extraction, all samples (Table S1) were stored in 100% ethanol at 4°C. Genomic DNA was extracted from samples using the DNeasy Blood & Tissue Kit (Qiagen) or the E.Z.N.A. Mollusc DNA Kit (Omega Bio-tek) (Table S2). High quality and high molecular weight genomic DNA (determined by gel electrophoresis) was diluted to 2.5 ng/µl and sequenced on an Illumina MiSeq (Illumina), using the Illumina Nextera XT library preparation reagents (Illumina). Paired-end, shotgun genomic DNA sequencing was performed using the Illumina MiSeq Reagent Kit v2/v3. MiMi analysis was conducted on eight individuals of each species (*P. miliaris* and *C. caeruleus*) which were indexed, pooled and sequenced on a flowcell, per species. For traditional microsatellite detection, single samples of each species (*T. eurycerus isaaci* and *N. pygmaeus*) were individually indexed, pooled and sequenced along with other species not used in this study (Table S2). Both methods were not tested for all species, due to these microsatellite markers being designed for active research projects that progressed beyond marker development as the MiMi method was being developed and iterated upon.

2.2 | MiMi microsatellite detection methodology

Microsatellite markers were initially designed in data from each sample using the pal_finder (Castoe et al., 2015) workflow of Griffiths et al. (2016); a traditional design method using the data of a single individual. A novel quality control procedure was developed for those data sets in which multiple individuals of the same species were sequenced (two species) with the aim of identifying polymorphic loci, filtering out primer pairs containing point mutations within the priming regions, and avoiding other potential issues with a locus including nonspecific primer binding and insertion/deletion mutations in the flanking regions. Eight individuals per species were sequenced and the data pertaining to each individual were first passed separately through the traditional design method. The eight individual output files then become the input for the novel method: Multi-individual Microsatellite identification (MiMi). MiMi takes the primer sequences developed in each individual and checks for their presence in the data of every other individual. Primer pairs for which the forward primer appeared in more than 33% of the individuals were selected and all reads containing the exact primer sequence compiled into an MSA file with the FASTA format. The MSA files were aligned using the MUSCLE alignment algorithm (Edgar, 2004) and putative loci automatically filtered to remove monomorphic loci, low quality "gapped" alignments and loci containing sequence mutations within the primer binding sites. Loci passing all filters are retained as high quality loci and loci passing some filters but lacking enough information to confidently pass all filters are retained as good quality loci. Both high quality and good quality loci are each ranked by the size range in alleles detected. A log file is produced detailing loci which have been removed by each filter. A Python script implementing the MIMI tool is available to download and run from https://github.com/graemefox/mimi.

2.3 | Optimisation of potential markers

Primer pairs developed under either design method were tested in 5 µl reactions using the Type-it Microsatellite PCR Kit (Qiagen) using the standard protocol and thermal cycling parameters (5 min at 95°C, 25–28*[30 s at 95°C, 90 s at 60°C, 30 s at 72°C], 30 min at 60°C). Only a single annealing temperature (60°C) was tested, as Primer3 (Koressaar & Remm, 2007; Untergasser et al., 2012) which is used during the traditional marker design process (Castoe et al., 2015; Griffiths et al., 2016), had been configured specifically for these PCR reagents and a primary goal of this method was to avoid time consuming annealing temperature optimisation. A marker was given successful amplification status if clean PCR products were clearly visible

TABLE 1 A summary of the design methods used in each species, including the data set number (ID), species, treatment (T_x), number of individuals sequenced (*N*), number of PCR primers tested (Pp), number of PCR primers tested successfully amplifying in 75% of samples tested (Amp), number of amplifiable PCR primers producing informative data after capillary electrophoresis (easily interpretable and polymorphic) (Inf), percentage of amplifiable primers which were informative (Inf/Amp), percentage of total primers tested which were informative (Inf/Pp) genome size estimate (C-val), raw sequence reads per sample (Reads), (mean and *SD* given where MiMi applied), estimated sequence coverage (Cov), literature reference and/or accession numbers of NGS data (REF/SRA) where applicable. All genome sizes were retrieved from the Animal Genome Size Database (www.genomesize.com) with the closest related species used. Panels of markers were developed in *P. miliaris* and *C. caeruleus* using both the traditional method (Castoe et al., 2015; Griffiths et al., 2016) and MiMi methods. The application of the MiMi quality control process produces higher rates of both amplification and production of informative markers in both these instances

ID	Species	T_x	*N*	Pp	Amp	Inf	Inf/ Amp	Inf/ Pp	C val	Reads	Cov	REF/ SRA
1	*C. caeruleus*	MiMi	8	10	10	8	80%	80%	1.47	8* 2,901,027, (STDEV ± 878, 838)	1.20X	SRX5066864 to SRX5066869
2	*P. miliaris*	MiMi	8	20	19	18	95%	90%	1.30	8* 1,482,736, (STDEV ± 280, 686)	0.57X	SRX5162614 to SRX5162621
3	*T. eurycerus isaaci*	Trad.	1	30	21	18	86%	60%	3.94	8,980,510	1.10X	Combe et al. (2018)/ SRX5116712
4	*N. pygamaeus*	Trad.	1	30	26	17	65%	57%	3.58	5,309,686	0.74X	SRX5112421
5	*C. caeruleus*	Trad.	1	10	4	1	25%	10%	1.47	3,913,299	1.60X	SRX5066867
6	*P. miliaris*	Trad.	1	24	13	9	69%	38%	1.30	1,359,615	0.52X	SRX5162614

on a 2% agarose gel in the 100–1,000 bp range for six or more individuals out of eight tested. Fluorescent dyes (6-FAM, TAMRA, HEX, PET) were added to PCR products using a universal tail technique (Blacket, Robin, Good, Lee, & Miller, 2012). Fragment length was determined using an ABI 3730 DNA Analyzer capillary sequencer (ThermoFisher Scientific) with GeneScan 500 LIZ dye Size Standard (ThermoFisher Scientific) and analysed using GENEMAPPER 5.0 software (ThermoFisher Scientific). We define an informative marker as one that produces clearly interpretable electropherogram traces after capillary electrophoresis and is polymorphic in terms of PCR fragment length between multiple individuals.

3 | RESULTS

Of the markers which passed each set of quality controls, we were able to optimise amplifiable and informative markers at a rate of 47.9% using the traditional design method, and 86.6% using MiMi. Comparisons between average rates of successful amplification and production of informative loci for each marker design method demonstrated a marked increase in both measures when MiMi was applied. In *P. miliaris* and *C. caeruleus*, markers were designed using both the traditional methodology and the MiMi methodology. A direct comparison between these two methods shows a very notable increase in both the rate of amplification success and the rate of development of informative markers (Figure 1). In two further species, (*T. eurycerus isaaci* and *N. pygmaeus*), markers were designed using only the traditional methodology. Rates of success for these species are presented as further evidence of a baseline of microsatellite design against which the MiMi method can be compared (Table 1). Unsuitable markers were removed at each filtering stage, reducing hundreds of thousands of possible markers designed by pal_finder, to a fewer than a hundred identified as high- or good-quality using MiMi (Table 2). Where MiMi was applied, the number of individuals sharing each common primer sequence ranged from three to seven (Figure 2). In the two example MiMi data sets presented here, 5% of potential loci were detected in sufficient individuals to allow further analysis by MiMi.

Automatic analysis of MSA files allowed the identification and removal of loci with mutations within the primer binding sites (Figures S1a,b) and loci showing very low alignment quality. Low alignment quality is indicative of a locus potentially containing fragment length altering polymorphisms (insertions/deletions) between the primer binding sites but outside the microsatellite locus itself (Figure S1c) or nonspecific primer binding. Monomorphic loci were also removed (Figures S1d,e). Of the markers which MiMi detected in multiple individuals, we were able to discount 79.3% of potential loci as unsuitable for microsatellite analysis (Table 3). High quality loci (those which exclusively showed evidence of positive characteristics) were detected at a rate of 4.5%, and good quality loci (those which did not show any evidence of negative characteristics, but did not have enough data to confidently pass all filters) were detected at an average rate of 16.1%.

TABLE 2 The total number of potential microsatellite loci discovered using the traditional design methodology, retained after filtering with the Griffiths et al. (2016) method and retained after MiMi quality control processing

Species	pal_finder loci	Griffiths et al. (2016) loci	MiMi loci
Cyanistes caeruleus	158,147	4,513 (2.9%)	302 (0.19%)
Psammechinus miliaris	469,047	5,657 (1.2%)	250 (0.05%)

Whilst the full MiMi method requires more data than the traditional approach detailed here (we recommend a minimum of eight individuals to be sequenced using the capacity of an entire MiSeq flowcell, although fewer samples are possible), the reduction in time spent in the laboratory, and associated savings, justifies the larger outlay in initial sequencing costs. A recent Illumina MiSeq run cost approximately $2,330, and using MiMi we recorded that 90% of the primer pairs chosen to be tested were successfully developed as informative microsatellite markers (Table 1, data set No.2). Using the traditional method, sequencing costs were less, as only a fraction (12.5%) of the capacity of a MiSeq sequencing flowcell was required, but only 38% of primer pairs tested were ultimately found to be informative markers (Table 1, data set No.5). The reduction in time and laboratory expense associated with investing in "failed markers" (inconsistent amplification/non polymorphic loci) ultimately results in a net saving when using MiMi. Based on our estimated rate of successful marker development, a project to develop a panel of 20 optimised markers over a two-week period using the MiMi methodology would cost less than using the traditional methodology over a four week period (16% reduction in total cost, 50% reduction in staff costs only, 19% increase in reagent costs only; see Tables S3 and S4). The most significant savings will be in researcher time spent screening loci, which was approximately 50% less using MiMi.

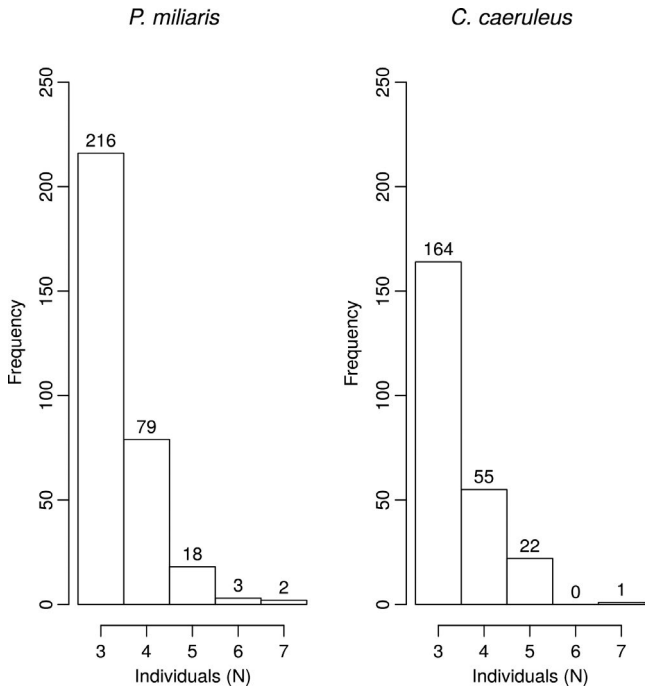


FIGURE 2 The MiMi tool was used to analyse 5,657 potential microsatellite loci discovered in *P. miliaris* sequence data and 4,513 discovered in *C. caeruleus*. Loci were filtered to just those which appeared in the sequence data of three or more individuals. The total number of loci which were successfully detected in multiple individuals, and in how many individuals they were detected is shown below. The bar labels are the absolute number of loci that were detected in each category (number of individuals)

3.1 | Description of output files

The outputs from the MiMi method are two tab separated tables containing details of the loci that have passed the quality control processes, a log file detailing which loci were removed under which quality-control conditions, and a per-locus MSA file in the FASTA format. The output tables each give the following information for each locus: forward primer sequence; reverse primer sequence; number of alleles at the locus; number of individuals in which the locus was sequenced in the data set; a description of the alleles found (the repeat motif and the number of repeats), and the predicted size range of amplicons produced using the PCR primers. The file "MiMi_output_all_loci.txt" gives details of every loci which MiMi was able to detect in multiple individuals (above the user-defined threshold) and "MiMi_output_filtered_loci.txt" gives just those loci which were able to pass all quality control filters as either high- or good quality. The log file details which loci were removed under which quality control conditions. Examples of the "MiMi_output_filtered_loci.txt" files resulting from the the MiMi analysis of *C. caeruleus* (data set No. 1) and *P. miliaris* (data set No. 2) are presented in Tables S5a,b, respectively. Three MSA files per locus are created: one containing the raw sequences from the input data that were found to contain the locus within the length of the read (ending ".fastq"); one containing these reads after alignment by MUSCLE (ending ".aln") and one containing aligned reads trimmed to the position of the forward primer (ending ".trimmed"). The main section of the MSA file name is the forward primer sequence of the locus.

4 | DISCUSSION

MiMi has proved to be a fast, cost effective approach to identification and characterisation of microsatellite markers using genomic sequence data from multiple individuals. The application of a microsatellite-picking tool such as pal_finder typically results in tens of thousands of potential loci, and therefore it makes logical sense to attempt to apply in silico marker optimisation methods over laboratory optimisation, to increase the efficiency in identifying informative loci. MiMi is the first tool, to our knowledge, that allows this range of

TABLE 3 Potential loci are automatically filtered by the MiMi script. Loci are removed under the following conditions: Low quality alignments = loci rejected due to not meeting a minimum requirement for overall quality of alignment. This is indicative of multiple primer binding occurring in the host genome, and of size-altering INDEL mutations occurring in the flanking regions. Primer mutations = loci rejected due to SNP or INDEL mutations detected within the primer binding sites. Nonvariable = loci rejected due to multiple reads spanning the microsatellite but no motif number variation present. High quality = loci passed due to consistent forward and reverse primer sequences seen in multiple individuals, multiple reads spanning the microsatellite and variable motif number observed, no evidence of INDEL or multiple binding sites, Good quality = identical criteria as "High quality," but alignment provided no information afforded relating to consistent reverse PCR primer or INDEL mutations

ID	Species	Total	Low quality alignments	Primer mutations	Nonvariable	High quality	Good quality
1	*Cyanistes caeruleus*	302	14 (4.6%)	7 (2.3%)	205 (67.9%)	13 (4.3%)	63 (20.9%)
2	*Psammechinus miliaris*	250	102 (40.8%)	9 (3.6%)	101 (40.4%)	12 (4.8%)	26 (10.4%)

important characteristics to be observed at the marker design stage (but see Nichols, Conroy, Kasinadhuni, Lamont, & Ogbourne, 2018). In a direct comparison between the traditional and MiMi methods, we show that the application of MiMi resulted in a 58% increase in the rate of identification of informative microsatellite markers, facilitating a 16% reduction in costs associated with the development of a microsatellite marker panel. To provide a baseline value of microsatellite design success, we also provide success rates for two species which only used the traditional methodology. Although not a true comparison, it appears that MiMi can be expected to produce amplifiable, informative markers at a consistently higher rate than the traditional methodology, facilitating an increase from ~57%–60% (data sets Nos. 3 and 4) to ~80%–90% (data sets Nos. 1 and 2). We feel certain that an increase of this order of magnitude, and the reduction in costs associated with the testing of markers which ultimately fail, fully justify the slight increase in sequencing costs associated with MiMi.

The incorporation of multiple genomes and construction of an MSA for each microsatellite locus allows several important quality checks to be made of each locus and facilitates notable increases in both the rate of successful amplification by PCR, and the development of informative markers. Nucleotide polymorphisms and INDEL mutations within the forward or reverse primer binding site can cause issues with inconsistent or failed PCR amplification, potentially resulting in allelic dropout (Silva, Torrezan, Brianese, Stabellini, & Carraro, 2017), and can also lead to an increase in the frequency of null alleles (Rico et al., 2017). Allelic dropout can present a significant problem during microsatellite analysis, causing decreased estimates of observed heterozygosity and increased estimates of inbreeding in the population (Wang, Schroeder, & Rosenberg, 2012). Two main causes of allelic dropout have been shown: sequence variation at a primer binding site (Silva et al., 2017) and PCR product size (particularly problematic for markers with large repeat counts; Sefc, Payne, & Sorenson, 2003). Through the construction of each MSA we were able to use MiMi to automatically confirm that primer-binding sites show strong sequence conservation, albeit in only a small subset of samples, thus minimising the likelihood that a putative marker would exhibit an elevated rate of allelic dropout caused by mis-priming. Confirmation of sequence conservation in at least one primer-binding site improved the rate at which we were able to amplify loci successfully. If possible, genomes of individuals from a range of putative populations should be included in the MiMi analysis to minimise null allele bias towards a particular sub population (Oosterhout, Weetman, & Hutchinson, 2005). Analysis of each microsatellite locus in an MSA also allows visualisation of the number of motif repeats, and automatic prioritisation of loci where variation is seen among samples. Rejecting monomorphic loci through MiMi produced an increase in the rate at which we were able to develop informative markers, compared to our own previous experience using other methods, and rates stated in the literature (Zhan et al., 2016). Additionally, MiMi automatically assesses the likelihood of the presence of multiple primer binding sites in the host genome by collating all sequences containing a common primer sequence. Where sequences containing the primer sequence produce low-overlap alignments, it is indicative that the corresponding primer binding site occurs in multiple locations across the genome, and thus that particular primer pair should be avoided to reduce cross-amplification.

Statistical models based upon a particular model of evolution at the microsatellite locus (the stepwise mutation model, for example) rely upon the assumption that the source of variation in fragment size is polymorphism in the number of repeats in the SSR (Dieringer & Schlötterer, 2003). The presence of other fragment length altering mutations between the primer binding sites (excluding the microsatellite itself) is indistinguishable by capillary electrophoresis from "true" variation at the microsatellite locus (Angers & Bernatchez, 1997; Grimaldi & Crouau-Roy, 1997; Stágel et al., 2009). Markers with fragment-length-altering mutations outside the microsatellite locus, potentially invalidate the assumptions of a number of models of microsatellite evolution, and are therefore avoided in our protocol.

Whilst MiMi does not allow one to state with certainty that a putative marker will not exhibit any of the negative characteristics described (allelic dropout, null alleles arising from population differentiation, nonvariable microsatellite loci, cross amplification or invalidation of assumptions of evolutionary model) when comprehensively characterised in a much larger number of samples, the opportunity to identify loci that do exhibit them, and subsequently remove them from analyses, is nevertheless valuable.

Variation in the rate at which loci were removed under each quality control category shows the importance of making each check, and that marker development in different taxa may perform differently from one another. In both examples of the application of MiMi here, we were able to remove undesirable loci that failed at least one quality check. Considering the total markers designed and filtered in both species, we were able to pass many loci (mean: 20.7%) that did not show evidence of these negative characteristics in the eight tested samples.

The success of MiMi is dependent upon the sequence coverage achieved in each sequencing run. Very low sequence coverage would probably result in relatively little overlap in the sequences of each individual, and therefore few loci passing the MiMi filter. The development of a new marker panel is very often performed in non-model species of specialised interest and it is likely that the genome size will be unknown and sequence coverage incalculable (Shikano, Ramadevi, Shimada, & Merilä, 2010). MiMi was successfully implemented in the two species tested here (with estimated coverage of 0.57X and 1.20X), suggesting that the method is suitable for genomic data sets with relatively low sequence coverage (Ekblom & Wolf, 2014). The proportion of individuals in which a primer must be detected is user definable, with a minimum of two individuals required for MiMi to provide useful information. Where loci were successfully detected in multiple individuals, we found a negative correlation between the number of potential markers and the frequency at which loci were found in multiple data sets. These frequencies are dependent upon the genome size, and the microsatellite richness of the genome, of the species of interest. Where estimates of genome coverage are approximately 1X or below, removal of duplicate primers/loci from the data set of each individual is recommended (implemented automatically in the Griffiths et al. (2016) workflow) as coverage of >1X of a locus in a single individual does not contribute any additional information to the MiMi process. However, where estimated coverage is significantly >1X, their removal may result in the dismissal of an increased frequency of otherwise useful loci that appear multiple times in the sequence data as a result of the random nature of shotgun sequencing (Bouck, Miller, Gorrell, Muzny, & Gibbs, 1998). In the event of a low number of markers ultimately being returned, the filter that removes loci appearing more than once in the data can easily be disabled at the web interface of the Griffiths et al. (2016) tool. In this case, multiple reads containing the primer sequence from the same biological sample will appear alongside each other in the output MSA, allowing the user to assess the reads as "shotgun duplicates" (i.e., multiple sequence reads covering the same genomic region of an individual, by chance).

MiMi makes several important assumptions of the characteristics of microsatellite loci investigated in a small number of samples, and infers these are representative of the loci in the wider population. However, this is not always expected to be true (Goldstein, Linares, Cavalli-Sforza, & Feldman, 1995) and the removal of otherwise useful markers, under the limiting assumptions of the MiMi quality control process, is likely to happen. For example, SSRs that do not show any variation in number of repeats in the sequence data are removed, but these loci may show variation in the wider population. The ethos behind the MiMi method is to select markers for which we have the most information, rather than seeking to discover as many markers as possible. Given the large numbers of potential markers we derived from the MiMi process, we do not consider the removal of potentially useful markers as a major disadvantage, and these markers can always be added back if needed.

Loci that do show allelic variation are ranked by the range size of the microsatellite repeat number (Goldstein & Schlötterer, 1999), with the assumption that the loci with the largest differences are most likely to be informative markers. A large range in the number of repeats implies that the variation seen at the locus is less likely to be the result of an amplification or sequencing error (Hosseinzadeh-Colagar, Haghighatnia, Amiri, Mohadjerani, & Tafrihi, 2016) but rather is representative of a true, variable microsatellite locus. We conclude that under the assumptions we identify here, the rate and efficiency of informative microsatellite discovery are greatly increased using high-throughput sequencing data in comparison to traditional microsatellite library discovery methods, but the robustness of MiMi should be tested in additional species.

We recommend that eight unrelated individuals are sequenced for MiMi processing for optimal capture of markers exhibiting multiple alleles at microsatellite loci. Whilst it is impossible to state an optimum figure for universal use, due to varying allelic richness in species and populations (Bashalkhanov, Pandey, & Rajora, 2009), in our experience, eight samples represents an acceptable balance between depth of sequencing coverage and allele rarefaction (Hale, Burg, & Steeves, 2012). In species where it is not feasible to source eight samples, related or not, due to their extreme scarcity, MiMi is still applicable. MiMi will function beneficially on any number of samples >1, whether related or unrelated. Furthermore, species with extremely large genomes may not perform well due to the limitations of sequencer capacity and the requirement for approximately 1X genome sequence coverage to be achieved. Our method has been tested on Illumina MiSeq data only, but will function on paired-end data, in the FASTQ format, from any sequencing platform, should additional depth of coverage be required. It is important to note that we are not attempting to detect all, or even most alleles present at a locus. Detecting the presence of multiple alleles (>1) is sufficient to enable MiMi processing. Other influencing factors, such as the sampling of related individuals or populations experiencing low genetic diversity due to historical population bottlenecks, may impact the allelic richness of the samples and therefore the ability of MiMi to detect multiple alleles (Price & Hadfield, 2014).

Methods of genotyping microsatellites by high-throughput sequencing are a promising development and avoid many of the ambiguities inherent in genotyping by capillary electrophoresis (Shin et al., 2017; Zhan et al., 2016). Determination of accurate genotypes by these methods enables many of the additional tests required of a

microsatellite marker (tests for linkage disequilibrium, frequency of null alleles, for example) to be carried out using NGS data alone. We envisage that large scale microsatellite studies be performed using two NGS runs: the first using MiMi to discover potentially informative microsatellites; and a second using a high-throughput genotyping method to genotype all experimental samples in one go (De Barba et al., 2016).

ACKNOWLEDGEMENTS

With thanks to the Genomic Technologies Core Facility of the University of Manchester for their expertise and services. *P. miliaris* samples were collected by Simon Exley of Queen's University Belfast. Funding for this PhD research comes from Manchester Metropolitan University. ACK thanks the Negaunee Foundation for their generous support of a curatorial preparator who extracted the samples of *N. pygmaeus*.

AUTHOR CONTRIBUTIONS

G.F., R.F.P., & J.K.R. conceived the concept. G.F. wrote and tested the programme and performed the marker optimisation in *C. caeruleus*. G.F., R.F.P., & J.K.R. verified the methods and the interpretation of the results. G.F., R.F.P., & J.K.R. discussed the results and drafted the manuscript, with helpful comments and contributions from remaining authors to the final manuscript. R.A. assisted with DNA sequencing at the University of Salford. F.C., and W.E.H. provided the *T. eurycerus isaaci* and *C. caeruleus* samples and associated sequence data, F.C. performed the marker optimisation in these species and F.C. ran the capillary sequencer. A.C.K., and A.N. provided the *N. pygamaeus* sample. M.B.S. provided the *P. miliaris* samples and sequence data, and performed the marker optimisation in this species. I.H. and S.R.D.K. provided the *C. cyanistes* samples.

DATA AVAILABILITY STATEMENT

The MiMi quality processing procedure is performed by an open-source Python script, freely available from https://github.com/graemefox/mimi. A small subset of example data is included at the repository. The pal_finder and pal_filter process required prior to MiMi is easily run and accessed via an online service hosted by the University of Manchester https://palfinder.ls.manchester.ac.uk/. Raw sequence reads are available from the N.C.B.I. BioProject and Sequence Read Archive (*C. caeruleus*: Accession PRJNA507250; *P. miliaris*: Accession PRJNA510714; *T. eurycerus isaaci*: Accession PRJNA509530; *N. pygamaeus*: Accession PRJNA509330).

ORCID

Graeme Fox https://orcid.org/0000-0001-7980-6944
Rachael E. Antwis https://orcid.org/0000-0002-8849-8194
Milena Benavides-Serrato https://orcid.org/0000-0002-1644-8673
Jennifer K. Rowntree https://orcid.org/0000-0001-8249-8057

REFERENCES

Abdelkrim, J., Robertson, B. C., Stanton, J. L., & Gemmell, N. J. (2009). Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *BioTechniques*, *46*(3), 185–192. https://doi.org/10.2144/000113084

Abdul-Muneer, P. M. (2014). Application of microsatellite markers in conservation genetics and fisheries management: Recent advances in population structure analysis and conservation strategies. *Genetics Research International*, *2014*, 1–11. https://doi.org/10.1155/2014/691759

Angers, B., & Bernatchez, L. (1997). Complex evolution of a salmonid microsatellite locus and its consequences in inferring allelic divergence from size information. *Molecular Biology and Evolution*, *14*(3), 230–238. https://doi.org/10.1093/oxfordjournals.molbev.a025759

Bashalkhanov, S., Pandey, M., & Rajora, O. P. (2009). A simple method for estimating genetic diversity in large populations from finite sample sizes. *BMC Genetics*, *10*(84), https://doi.org/10.1186/1471-2156-10-84

Bhargava, A., & Fuentes, F. F. (2010). Mutational dynamics of microsatellites. *Molecular Biotechnology*, *44*(3), 250–266. https://doi.org/10.1007/s12033-009-9230-4

Blacket, M. J., Robin, C., Good, R. T., Lee, S. F., & Miller, A. D. (2012). Universal primers for fluorescent labelling of PCR fragments–an efficient and cost-effective approach to genotyping by fluorescence. *Molecular Ecology Resources*, *12*(3), 456–463. https://doi.org/10.1111/j.1755-0998.2011.03104.x

Bloor, P. A., Barker, F. S., Watts, P. C., Noyes, H. A., & Kemp, S. J. (2001). *Microsatellite libraries by enrichment*. [online] Retrieved from http://www.genomics.liv.ac.uk/animal/MICROSAT.PDF

Bouck, J., Miller, W., Gorrell, J. H., Muzny, D., & Gibbs, R. A. (1998). Analysis of the quality and utility of random shotgun sequencing at low redundancies. *Genome Research*, *8*(10), 1074–1084. https://doi.org/10.1101/gr.8.10.1074

Castoe, T. A., Poole, A. W., de Koning, A. P. J., Jones, K. L., Tomback, D. F., Oyler-McCance, S. J., … Pollock, D. D. (2015). Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PLoS ONE*, *10*(8), e0136465.

Combe, F. J., Taylor-Cox, E., Fox, G., Sandri, T., Davis, N., Jones, M. J., … Harris, W. E. (2018). Rapid isolation and characterization of microsatellites in the critically endangered Mountain Bongo (*Tragelaphus eurycerus isaaci*). *Journal of Genetics*, *97*(2), 549–553. https://doi.org/10.1007/s12041-018-0922-z

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*, 499–510. https://doi.org/10.1038/nrg3012

De Barba, M., Miquel, C., Lobréaux, S., Quenette, P. Y., Swenson, J. E., & Taberlet, P. (2016). High-throughput microsatellite genotyping in ecology: Improved accuracy, efficiency, standardization and success with low-quality and degraded DNA. *Molecular Ecology Resources*, *17*(3), 492–507. https://doi.org/10.1111/1755-0998.12594

Dieringer, D., & Schlötterer, C. (2003). Two distinct modes of microsatellite mutation processes: Evidence from the complete genomic sequences of nine species. *Genome Resources*, *13*(10), 2242–2251. https://doi.org/10.1101/gr.1416703

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Ekblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*(9), 1026–1042. https://doi.org/10.1111/eva.12178

Fox, G., Darolti, I., Hibbitt, J. D., Preziosi, R. F., Fitzpatrick, J. L., & Rowntree, J. K. (2018). Genetic assessment of ex situ populations to aid species conservation and maintain heterozygosity in non-model

species. *Journal of Zoo and Aquarium Research*, *6*(2), 50–56. https://doi.org/10.19227/jzar.v6i2.299
Goldstein, D. B., Linares, A. R., Cavalli-Sforza, L. L., & Feldman, M. W. (1995). An evaluation of genetic distances for use with microsatellite loci. *Genetics*, *139*(1), 463–471.
Goldstein, D. B., & Pollock, D. D. (1997). Launching microsatellites: A review of mutation processes and methods of phylogenetic inference. *Journal of Heredity*, *88*(5), 335–342. https://doi.org/10.1093/oxfordjournals.jhered.a023114
Goldstein, D. B., & Schlötterer, C. (1999). *Microsatellites: Evolution & applications*. Oxford, UK: Oxford University Press.
Griffiths, S. M., Fox, G., Briggs, P. J., Donaldson, I. J., Hood, S., Richardson, P., … Preziosi, R. F. (2016). A Galaxy-based bioinformatics pipeline for optimised, streamlined microsatellite development from Illumina next-generation sequencing data. *Conservation Genetics Resources*, *8*(4), 481–486. https://doi.org/10.1007/s12686-016-0570-7
Grimaldi, M. C., & Crouau-Roy, B. (1997). Microsatellite allelic homoplasy due to variable flanking sequences. *Journal of Molecular Evolution*, *44*(3), 336–340. https://doi.org/10.1007/PL00006151
Hale, M. L., Burg, T. M., & Steeves, T. E. (2012). Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. *PLoS ONE*, *7*(9), e45170. https://doi.org/10.1371/journal.pone.0045170
Hosseinzadeh-Colagar, A., Haghighatnia, M. J., Amiri, Z., Mohadjerani, M., & Tafrihi, M. (2016). Microsatellite (SSR) amplification by PCR usually led to polymorphic bands: Evidence which shows replication slippage occurs in extend or nascent DNA strands. *Molecular Biology Research Communications*, *5*(3), 167–174.
Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, *155*(1), 27–38. https://doi.org/10.1016/j.cell.2013.09.006
Koressaar, T., & Remm, M. (2007). Enhancements and modifications of primer design program Primer3. *Bioinformatics*, *23*(10), 1289–1291. https://doi.org/10.1093/bioinformatics/btm091
McPherson, J. D. (2014). A defining decade in DNA sequencing. *Nature Methods*, *11*, 1003–1005. https://doi.org/10.1038/nmeth.3106
Nichols, J., Conroy, G. C., Kasinadhuni, N., Lomont, R. W., & Ogbourne, S. M. (2018). In silico detection of polymorphic microsatellites in the endangered Isis tamarind, Alectryon ramiflorus (Sapindaceae). *Applications in Plant Sciences*, *6*(11), e01196. https://doi.org/10.1002/aps3.1196
Oosterhout, C. V., Weetman, D., & Hutchinson, W. F. (2005). Estimation and adjustment of microsatellite null alleles in nonequilibrium populations. *Molecular Ecology Notes*, *6*(1), 255–256. https://doi.org/10.1111/j.1471-8286.2005.01082.x
Price, M. R., & Hadfield, M. G. (2014). Population genetics and the effects of a severe bottleneck in an ex situ population of critically endangered Hawaiian tree snails. *PLoS ONE*, *9*, e114377. https://doi.org/10.1371/journal.pone.0114377
Puckett, E. E. (2016). Variability in total project and per sample genotyping costs under varying study designs including with microsatellites or SNPs to answer conservation genetic questions. *Conservation Genetics Resources*, *9*(2), 289–304. https://doi.org/10.1007/s12686-016-0643-7
Ribout, C., Villers, A., Ruault, S., Bretagnolle, V., Picard, D., Monceau, K., & Gauffre, B. (2019). Fine-scale genetic structure in a high dispersal capacity raptor, the Montagu's harrier (*Circus pygargus*), revealed by a set of novel microsatellite loci. *Genetica*, *147*(1), 69–78. https://doi.org/10.1007/s10709-019-00053-7
Rico, C., Cuesta, J. A., Drake, P., Macpherson, E., Bernatchez, L., & Marie, A. D. (2017). Null alleles are ubiquitous at microsatellite loci in the Wedge Clam (*Donax trunculus*). *PeerJ*, *5*, e3188. https://doi.org/10.7717/peerj.3188
Rose, O., & Falush, D. (1998). A threshold size for microsatellite expansion. *Molecular Biology and Evolution*, *15*(5), 613–615. https://doi.org/10.1093/oxfordjournals.molbev.a025964
Sefc, K. M., Payne, R. B., & Sorenson, M. D. (2003). Microsatellite amplification from museum feather samples: Effects of fragment size and template concentration on genotyping errors. *The Auk*, *120*(4), 982–989. https://doi.org/10.1642/0004-8038(2003)120[0982:MAFMFS]2.0.CO;2
Shikano, T., Ramadevi, J., Shimada, Y., & Merilä, J. (2010). Utility of sequenced genomes for microsatellite marker development in non-model organisms: A case study of functionally important genes in nine-spined sticklebacks (*Pungitius pungitius*). *BMC Genomics*, *11*(334), https://doi.org/10.1186/1471-2164-11-334
Shin, G., Grimes, S. M., Lee, H. J., Lau, B. T., Xia, L. C., & Ji, H. P. (2017). CRISPR-Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis. *Nature Communications*, *8*(14291). https://doi.org/10.1038/ncomms14291
Silva, F. C., Torrezan, G. T., Brianese, R. C., Stabellini, R., & Carraro, D. M. (2017). Pitfalls in genetic testing: A case of a SNP in primer-annealing region leading to allele dropout in *BRCA1*. *Molecular Genetics and Genomic Medicine*, *5*(4), 443–447. https://doi.org/10.1002/mgg3.29
Stágel, A., Gyurján, I., Sasvári, Z., Lanteri, S., Ganal, M., & Nagy, I. (2009). Patterns of molecular evolution of microsatellite loci in pepper (*Capsicum* spp.) revealed by allele sequencing. *Plant Systematics and Evolution*, *281*(1–4), 251–254. https://doi.org/10.1007/s00606-009-0196-2
Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3–new capabilities and interfaces. *Nucleic Acids Research*, *40*(15), e115. https://doi.org/10.1093/nar/gks596
Vieira, M. L. C., Santini, L., Diniz, A. L., & Munhoz, C. F. (2016). Microsatellite markers: What they mean and why they are so useful. *Genetics and Molecular Biology*, *39*(3), 312–328. https://doi.org/10.1590/1678-4685-GMB-2016-0027
Wang, C., Schroeder, K. B., & Rosenberg, N. A. (2012). A maximum-likelihood method to correct for allelic dropout in microsatellite data with no replicate genotypes. *Genetics*, *192*(2), 651–669. https://doi.org/10.1534/genetics.112.139519
Witzenberger, K. A., & Hochkirch, A. (2011). Ex situ conservation genetics: A review of molecular studies on the genetic consequences of captive breeding programmes for endangered animal species. *Biodiversity and Conservation*, *20*(9), 1843–1861. https://doi.org/10.1007/s10531-011-0074-4
Zhan, L., Paterson, I. G., Fraser, B. A., Watson, B., Bradbury, I. R., Ravindran, P. N., … Bentzen, P. (2016). MEGASAT: Automated inference if microsatellite genotypes from sequence data. *Molecular Ecology Resources*, *17*(2), 247–256. https://doi.org/10.1111/1755-0998.12561

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Fox G, Preziosi RF, Antwis RE, et al. Multi-individual microsatellite identification: A multiple genome approach to microsatellite design (MiMi). *Mol Ecol Resour.* 2019;00:1–9. https://doi.org/10.1111/1755-0998.13065

7.3 Appendix 3 - Supplementary information relating to Chapter 3.

Supplementary information published in Molecular Ecology Resources relating to Chapter 3.

Supplementary Information for:

Multi-individual Microsatellite identification: a multiple genome approach to microsatellite design (MiMi).

Graeme Fox[1], Richard F. Preziosi[1], Rachael E. Antwis[2], Milena Benavides-Serrato[1, 3],Fraser J. Combe[4], W. Edwin Harris[5], Ian R. Hartley[6], Andrew C. Kitchener[7], Selvino R. de Kort[1], Anne-Isola Nekaris[8], Jennifer K. Rowntree[1]*

[1] Ecology and Environment Research Centre, Department of Natural Sciences, Manchester Metropolitan University, Manchester, M1 5GD, UK.

[2] School of Environment and Life Sciences, University of Salford, Salford, M5 4WT, UK.

[3] Universidad Nacional de Colombia, Sede Caribe-CECIMAR Calle 25 #2-55, Playa Salguero, Colombia.

[4] Kansas State University, Division of Biology, Manhattan, KS, United States.

[5] Crop and Environment Sciences, Harper Adams University, Newport, TF10 8NB, UK.

[6] Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK.

[7] Department of Natural Sciences, National Museums Scotland, Chambers Street, Edinburgh, EH1 1JF, UK.

[8] Department of Social Sciences, Faculty of Humanities and Social Sciences, Oxford Brookes University, Oxford, OX3 0BP, UK.

*Correspondence: j.rowntree@mmu.ac.uk

Running title: Multi-individual microsatellite ID

Table S1.
Details of samples used in this study, where known or applicable, including species name, researcher responsible for sample provision (all sample providers are co-authors and can be identified by their initials), number of samples sequenced, co-ordinates of sample sites, relatedness, sex, details of any ethical or licensing requirements.

Species	Sample Provider	Sample Size	Sample Sites (co-ords)	Relatedness	Sex	Ethical/Licensing Details
Cyanistes caeruleus	IRH, SRdK	8	All Lancaster University. N 54.010933, W 2.787346	1) None 2) None 3) Father of 5 4) Mother of 5 5) Offspring of 3 & 4 6) Sibling of 5 7) Sibling of 5 8) None	1) Male 2) Female 3) Male 4) Female 5) Unknown 6) Unknown 7) Unknown 8) Unknown	All blood samples were collected under Home Office and Natural England licences to Ian Hartley and sampling protocols were approved by the Lancaster University Animal Welfare & Ethical Review Body.
Psammechinus miliaris	MBS	8	Fraoch Eilean. N 56.05975, W 5.14511	Unrelated	Unknown	None
Tragelaphus eurycerus	FC, WEH	1	Captive	Unrelated	Unknown	Samples collected from UK zoo animals. No ethical or licensing information required.
Nycticebus pygamaeus	AN, ACK	1	Captive	Unrelated	Unknown	Samples collected from EU zoo animals. No ethical or licensing information required.

Table S2.
Details of the dataset ID, species, DNA extraction reagents, MiSeq sequencing reagents version, sequencing read length, approximate proportion of flowcell and sequencing centre used to process and sequence samples for each dataset. Where samples from multiple species were sequenced on a single flowcell, the approximate proportion of flowcell capacity is given as a percentage. For example, a single sample of *T. eurycerus* was sequenced on a flowcell containing another single sample from a different species. In this case *T. eurycerus* utilised approximately 50% of the flowcell capacity. Eight *C. caeruleus* samples were sequenced on a flowcell, with no other samples. In this instance, *C. caeruleus* utilised 100% of the flowcell capacity.

Dataset ID	Species	DNA Extraction Reagents	MiSeq Sequencing Reagents	Mean Read Length (bp)	% of Flow Cell	Sequencing Facility
#1 / #5	*Cyanistes caeruleus*	DNeasy Blood and Tissue (Qiagen)	v3	300	100	University of Salford, School of Environment and Life Sciences
#2 / #6	*Psammechinus miliaris*	E.Z.N.A Mollusc DNA (Omega bio-tek)	v2	250	100	University of Manchester Genomic Technologies Facility
#3	*Tragelaphus eurycerus*	DNeasy Blood and Tissue (Qiagen)	v2	250	~50	University of Manchester Genomic Technologies Facility
#4	*Nycticebus pygamaeus*	DNeasy Blood and Tissue (Qiagen)	v2	250	~33.3	University of Salford, School of Environment and Life Sciences

Table S3.
Additional information regarding typical estimated reagent costs relating to developing a microsatellite marker panel. Prices stated are list prices as in June 2018.

DNA Extraction	List Price	
Isolate II Genomic DNA Kit (Bioline)	£149 (50 rx)	£2.98 per extraction
PCR Primer Synthesis (pair) (Sigma Aldrich)	£10 approx.	£10 per pair of markers tested
Type-it Microsatellite PCR Kit (Qiagen)	£1291 (2000rx)	£0.13 per 5µL reaction (reaction volume reduced to 20%)
Hi-Di Formamide (ThermoFisher Scientific)	£45.90	£1.60 per 96 reactions
GeneScan 500 LIZ dye Size Standard	£584	£7 per 96 reactions (volume reduced to 10%)
Capillary Electrophoresis	£1 per lane	£96 per 96 lanes

Table S4.
Breakdown of estimated cost of microsatellite panel development, June 2018. At a 90% success rate (MiMi method), to develop 20 markers 23 primers pairs require testing. At a 28% success rate (traditional method), to develop 20 markers 72 primer pairs require testing. Using the traditional method less than half a flowcell is required and the remaining capacity is used by another sample, sequenced on the same flowcell, but for a different experiment. Using the MiMi method the capacity of an entire flowcell is required to get sufficient coverage for the eight samples sequenced. Prices are estimates.

	MiMi (90% Successful Development Rate)	Traditional (28% Successful Development Rate)
Illumina MiSeq Sequencing	~£1800 (1 flowcell)	~£900 (0.5 flowcell)
DNA Extraction (8 samples)	£23.84	£23.84
PCR Primer Synthesis	~£230 (23 primer pairs)	~£720 (72 primer pairs)
'Type-it' PCR Reagents	~£45 (500rx)	~£90 (1000rx)
Capillary Electrophoresis + Reagents	~£200	~£200
Technicians Salary (£25K Per annum)	~2 weeks (£961)	~4 weeks (£1923)
Total Estimate	£3259.84	£3856.84

Tables S5(a) Example MiMi output from the analysis of *C. caeruleus* (dataset #1), and Table S5(b) example output from the analysis of *P. miliaris* (dataset #2). Columns in both tables show the forward primer sequence, the reverse primer sequence, the number of unique alleles (variation in the number of repeats) detected, the number of individuals in which the microsatellite locus was detected, the alleles present (motif and total length of repeat region in nucleotides) and the size range (difference in nucleotides between the smallest and largest allele detected.

Table S5(a)

Forward Primer Sequence (5′-3′)	Reverse Primer Sequence (5′-3′)	Number of Alleles	Found in Individuals	Alleles Present	Size Range
AGGAAGGGACCAGACAATCC	GCATTTTCTCAATGTCAGACCC	2	3	ATCC(32) ATCC(60)	28
TGGGACAGGGAAAAGAAAGG	GTGATGGATGTGGCTGTTCC	2	4	TG(20) TG(36)	16
GCTGTGCTGAAGTTCCTTCG	CGCACATCTTGTGAAATTCG	2	3	TG(28) TG(22)	6
TGTCCCGTACACTGGAAAGG	TGGTTACCCAGTTTCTACTGCC	2	5	AG(12) AG(18)	6
CAACAGTTTTCTCTAGGCTGTGG	AATGGAGGGATTTCAGACAGC	3	4	AAAC(24) AAAC(28) AAAC(24)	4
ACAGAAGCCATGACAAGGGG	CCGGTACATGACATAGAGACTCACC	3	4	AG(20) AG(16) AG(16)	4
CATGGGACGTGAAGAGTATGG	TTCAGAGCCTTCATCACATCC	2	3	AG(24) AG(20)	4
ACCAGGTAGCTGTCAGTTGAGG	GATGTGAATGGACCAGATTGC	4	4	TG(12) TG(16) TG(16) TG(16)	4
ATAGGCCATGATCCCTTTCC	ACCTCAGCTTTGCTTTGTGG	2	5	TGC(27) TGC(30)	3
CTATGTGTGTGGCCAGTTGC	CCTGCACTCCAGATACCAGC	3	3	TG(16) TG(14) TG(14)	2
CTTGCTCTCTTCATCCTTCCC	GGCGTTGAAGTAGCTGAAGG	2	4	AG(12) AG(14)	2
TCCCTTCAGCACTGTCTGC	CCTATTTGTGTGTGTGGCCC	3	4	TG(14) TG(14) TG(12)	2
GGCAGAAGCATGTGAAATCC	GAGTGAACCAGCCACAGTCC	2	3	AG(16) AG(18)	2
CATTGTGAAAGGAGATGCCC	GGTTTGTATTCCCATCGTGC	2	3	AAAAGG(126) AAAAGG(54)	72
AAGCAGATAGCACTGGCAGC	CCATAACTCTCAACAGCCAAGC	2	3	ACT(60) ACT(18)	42
GTGGTTAAACCCCAACCAGC	ACCATTGCTTGGGAACAGG	2	3	ACTC(56) ACTC(24)	32
GAACACTGAATGTCTCTTCCAGC	CTGCTGTCACTCACCTGTGG	2	4	AATAG(60) AATAG(90)	30
CTACTCCAGGCTGAGGCTCC	GAATTGCTGCCTCCTCCC	2	3	TG(14) TG(38)	24
GCGATTAAGCCATCAATCTCC	GCTATAAAGTGCTGGAGCGG	2	4	TTG(39) TTG(18)	21
TCCAACAAATCCTGGAGTGC	CTTGCTCTGAAGCCTAGGGG	2	3	TGC(48) TGC(27)	21

TCAGGACATCTGTGAGCAGC	CCCTGCAAGGCTAAATCCC	3	3 ATAG(60) ATAG(40) ATAG(56)	20
TGCTGTTCTGAAGCAGTTGG	ACTGGGAGGTAAATTTGGGG	2	3 ATC(42) ATC(24)	18
TGGGAGGAAAATATGGGTGC	ATCCAAACTGTACATGCGCC	2	3 AG(12) AG(30)	18
GTGAGGCACCACTGGAAGC	TCTCCATTTGGCAATCTGTAGG	3	3 TTG(36) TTG(27) TTG(18)	18
TGCACCCTTTCACCTAGACC	GCTCTGTTCCCAGGATTGG	2	3 ATGGAG(36) ATGGAG(54)	18
AGATCCAACGGAGAGTGGG	CTTGGAGCAGTGATTTCAAGC	2	3 ATC(27) ATC(42)	15
CAGGGCTCTGAAGAAACTGC	GGCTGGTAGAGATGTGCAGG	2	4 ATCC(48) ATCC(36)	12
CACTGCAATGATTAAGGCTGC	GCCTAGCAGGATGAGATGGG	2	3 ATGAG(60) ATGAG(50)	10
GACCAGCTTTTCTCTTCCCC	CTGCACTAGGGAGCTGATGG	2	3 TG(22) TG(32)	10
AAACTGGCTTGTGTGAAGGG	TTAGGGAAACTGCAGCAAGG	2	3 AG(12) AG(22)	10
GGACAGGGATGCTAACAGGG	ATGCTGCTACAGCCAGCCC	3	3 ATC(30) ATC(39) ATC(30)	9
ATGGATTCGTTGCATTCAGG	TAAAGTCACCTGACCCTGCC	3	3 ATCC(52) ATCC(48) ATCC(44)	8
CTAGCTGCTGCCATAGGAGG	AGGAGTGTCTGCATTCCTGG	3	3 TG(24) TG(16) TG(18)	8
GGCACCAGATGCAGTAATATTGG	AGGCAAAGAGAACAGAGCAGC	2	3 AT(20) AT(12)	8
TGTGTCCTTAAAGCTAGGGGC	ACATTTAAGGGAGGTTGTGGC	2	3 ATAG(44) ATAG(36)	8
CAGGCCTTTGATAAGGTCCC	CTCTGGACAACATCCCATCC	2	3 ATCC(40) ATCC(48)	8
GGCAGGAGGACAAAAGAAGC	CATCCCTGAATTTCCAGCC	3	4 TG(18) TG(24) TG(16)	8
CAAGTGTTATGTGATAGAGGAGGGG	GCAGGTTCAGCATTTGTGG	2	3 AG(18) AG(24)	6
TTGGAAGGAGTTTCCAATGC	GTGTATGTGAGGATGTTAGCAAGC	2	3 ATTCTT(78) ATTCTT(72)	6
AGAACAGCAGCGTGAGTGC	CTGACCGCACAGAGACACC	2	3 AGG(18) AGG(24)	6
CTTGGCTGTAGCATTCTGGC	AGTCCAGTCACTTGGCATCC	2	4 AG(16) AG(22)	6
TGTTGAAGAGGCATTGCTGC	TTTCATCACCAGATGTCCCC	2	4 AGCTC(35) AGCTC(30)	5
TTCTTGCCTTTTGGAGATGC	TCCCCAGCTATTTGCTTACG	2	3 TG(20) TG(24)	4
CCGTATGTTTCTTAGGCCCC	ACAACCTGTTTGTGCAAGGC	3	3 AG(22) AG(22) AG(18)	4
CTTTCATTTCCCCTCCTCCC	CCAGATCAGGGTCACAGAGC	2	3 AG(16) AG(12)	4
AGATCCATGGAAGTAGGGAAGG	GATTGGAGAGGTGGGTTGG	3	3 TG(12) TG(16) TG(16)	4
GAATCCTGTTGCATTGAGCC	CGTCCTTCAGGACTGTCACC	2	3 TG(12) TG(16)	4
GCGATGAGTGGATTTTCTGC	GAAGGGGTGTTTGTTCCTGG	2	3 AG(12) AG(16)	4
CATGGCACTGACAGATTTTCC	TTTCAGAGGCACAAAGGACC	2	3 TG(12) TG(16)	4
CAGGCTGGGTTTAGGTTTGG	GGGTGAGCTCTGTATGCACC	2	3 AG(12) AG(16)	4

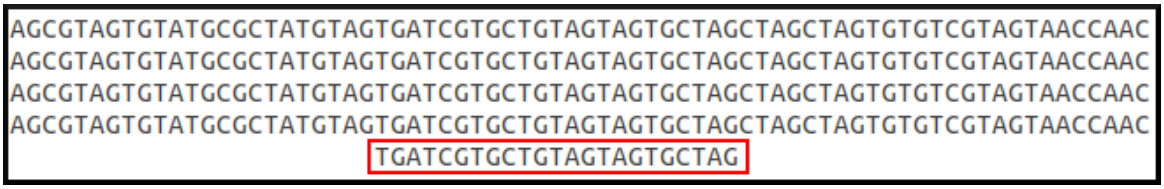
Forward Primer Sequence (5′-3′)	Reverse Primer Sequence (5′-3′)	Number of Alleles	Found in Individuals	Alleles Present	Size Range
TTTCACCACTCTCCTTCTCTCC	GTTCTCAAGCAGACGATGGG	3	4	TC(30) TC(16) TC(14)	16
GTGTTTGTGGGAGAGAAGGG	ATGAGTGTTCCCAAGGTTCG	2	3	TC(62) TC(46)	16
CCTCTGCTCACATACAGAGTCG	CTTTTATACCTCCGAGGCCC	3	3	TC(20) TC(14) TC(16)	6
AACTTTGGAGCAACGAAACG	TCAGTTGGTTCTATGCCTCG	2	3	ATG(30) ATG(24)	6
CGTGCGTACACAACACTTGC	CCTATCCTTCATGTCGGGC	3	4	TC(18) TC(22) TC(18)	4
TGAAATGTAGGAGGATGGGG	TCATAATGTGCATGCTTGGC	2	3	TC(16) TC(12)	4
TCAGTGAAGGAAGAAAGGCG	CTGTACGTGATTGCTGTCGG	2	4	TTC(21) TTC(18)	3
TCACTGCCACTGAAATTTGG	AACTTTGGAGCAACGAAACG	2	3	ATG(27) ATG(24)	3
CAGATTCAGAGTGATTGTGTGC	AACACCCACGAAAGGACC	2	3	AG(14) AG(16)	2
CGGAAGAGACCCTTTAAGTCAAATGAGG	CTCCCTGCCTGTTTACATCACTTCC	2	5	AG(16) AG(14)	2
TAGTCAATAAAGCGCAGCCC	TATCATGACCCTAGTGGCCG	3	3	AT(14) AT(12) AT(12)	2
GTCTCTTTCCGTCTCTCCCG	TGGATTGAGTTACCGCTTCG	2	3	TC(14) TC(12)	2
GACAGAGGGCAGTTATGATAAGG	GAAATTCGCTGGTGAAAACG	2	3	TTC(24) TTC(66)	42

GCCAGGAAAGTTCAATGTTGATAGCG	CACCCGCACATGAGCATCC	2	4	AG(44) AG(18)	26
CCAACTCTTTGTCACTGGGG	TGTGGCCTCAATGGAGTAGC	2	3	TC(28) TC(14)	14
TGCCTGTCTGTTTTGTGACG	AAGGGTTGAGCGAATGAGG	2	3	TC(32) TC(42)	10
TACTTTGCAAGGGTCAAACG	TCGCCAAAGTGCTAACTCG	3	3	TC(12) TC(22) TC(12)	10
CAGCACCTAATTATTCCCGC	AAGGGGAATAGGGGAATGG	3	5	TC(32) TC(28) TC(22)	10
TCATTGGGTCCTGATAAACTCC	GCTCTCAAGACATCCTTGCC	3	3	AC(12) AC(16) AC(20)	8
CGTTTAGACATCTTTCAGAGGACG	CCTTGGCTATAGGAGACCGC	2	3	TC(14) TC(22)	8
GCCTACTATCGACTCATTTTACTGGG	GTTATTACAAAGGTCGGGTTACCG	2	3	AT(20) AT(14)	6
GTCTCGAACGGAAGTTCAGG	ATTCATTACATGCAGCACGG	2	4	TC(32) TC(28)	4
TAATGGTGTCATGTCTCGGC	CAGTGATGATTTGGCTGGC	2	4	TC(28) TC(32)	4
CTGTCGCCTCCTTTTAATATGC	ATGGAGAGGAAAGCTGTTGG	2	3	TC(36) TC(32)	4
TGTGATATTTTGGTGAGCCG	TTTTGTGCTGGTTCGTGG	2	3	TC(16) TC(12)	4
CGATTTCTGATTACGCTTGC	GCGAGTGCAGTCTCTACGC	2	3	TTC(18) TTC(21)	3
GTGGTAACTATAATAGGGGCATGG	TTAAGGTGCATCCAGGTACG	2	3	ATG(48) ATG(45)	3
CGATACGGAAGCTAACAAACC	GCAAAAGGCCTTCAATAAGC	2	3	ATG(18) ATG(21)	3

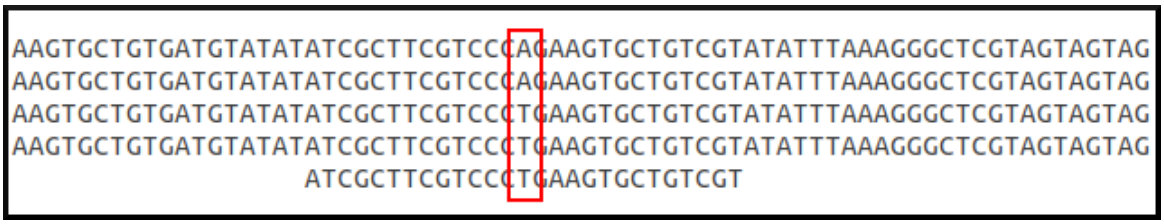
TGCTGTTGAATACCATTGCG	GCCCATCTCCACAACAGC	3	3	TGC(21) TGC(18) TGC(21)	3
TATTAGTTTGCGCAGGTTCG	TCAAACAAAGGATGAAGGGG	2	3	TC(20) TC(22)	2
TTATGAGCACCGGTCTAACG	GTACATGGCTCCAAGCAAGG	2	3	TC(20) TC(18)	2
ACGTGAGAATCAAAGCCCC	TATTTACCTTGCCCGAATGC	2	3	AT(14) AT(12)	2
CACCTCAAGTTTGCAATCCC	TTCAACCGCCTGGTTTAGC	2	3	TC(18) TC(16)	2
CCAAATCATAGGATGGTGGC	TCGGAAACTTTCACTCCTGC	2	3	TC(28) TC(30)	2
GGGTTGTTTGCTTGTTCGG	GCAGCTGAACTGAAGGTGG	4	4	TC(14) TC(16) TC(16) TC(16)	2
AAACTGTCAAGGAAGGCTGG	TGAGTGATGGTAGTTTCGCC	2	3	ATG(33) ATG(27)	6
GTGGGTGTCCTCATCAAACC	GTTGGTTTCAATACCACGGC	3	4	TGG(24) TGG(21) TGG(24)	3
TTTGAGAGAGATCGAACGGG	TTTCCCAAAGTCTCTCTGTGG	2	3	TC(22) TC(24)	2

Figure S1. Multiple sequence alignments showing benefits of a multi-individual approach to marker development.
(a) A primer sequence (highlighted in red) aligned against data from multiple samples. The primer is completely conserved in the four sequences above the primer. This primer will likely perform well in PCR due to the high rate of conservation and will likely not suffer an elevated rate of null alleles.
(b) A primer sequence aligned against data from multiple samples. The primer region contains a single nucleotide polymorphism at position 14 in the primer (highlighted in red). The top two template sequences contain adenine nucleotides and the lower two contain thymine. This primer may still perform generally well in PCR but is more likely to suffer from null alleles.
(c). A section of the flanking region between a primer binding site and the microsatellite region. Generally, the sequences are very highly conserved, however, here one sequence contains a significant deletion mutation (highlighted in red) whilst maintaining high sequence conservation at either side of the deletion. The middle sequence is missing 33 nucleotides compared to the other four sequences. Using traditional microsatellite genotyping methods which utilise molecular weight, the change in fragment size due to this deletion is undistinguishable from "true" mutations at the microsatellite, leading to a potentially incorrect genotype.
(d). A microsatellite region (highlighted in red: CATA motif) flanked by highly conserved sequence. The length of the SSR is the same in all five sequences (CATA*7), indicating that this particular marker may not be informative, as it does not show any fragment length altering polymorphisms within the microsatellite.
(e) A microsatellite locus with variation in the number of CATA repeats present. Five sequences are presented showing four different genotypes (highlighted in red) and maintenance of the high sequence conservation in the flanking regions at each side of the microsatellite. This locus is much more likely to be informative for genotyping as it shows the variation in the locus that is crucial for utility as a genetic marker.

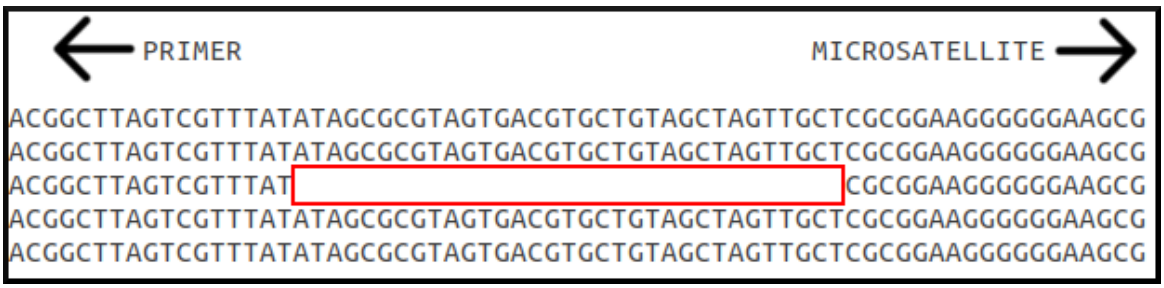
a)


AGCGTAGTGTATGCGCTATGTAGTGATCGTGCTGTAGTAGTGCTAGCTAGCTAGTGTGTCGTAGTAACCAAC
AGCGTAGTGTATGCGCTATGTAGTGATCGTGCTGTAGTAGTGCTAGCTAGCTAGTGTGTCGTAGTAACCAAC
AGCGTAGTGTATGCGCTATGTAGTGATCGTGCTGTAGTAGTGCTAGCTAGCTAGTGTGTCGTAGTAACCAAC
AGCGTAGTGTATGCGCTATGTAGTGATCGTGCTGTAGTAGTGCTAGCTAGCTAGTGTGTCGTAGTAACCAAC
TGATCGTGCTGTAGTAGTGCTAG

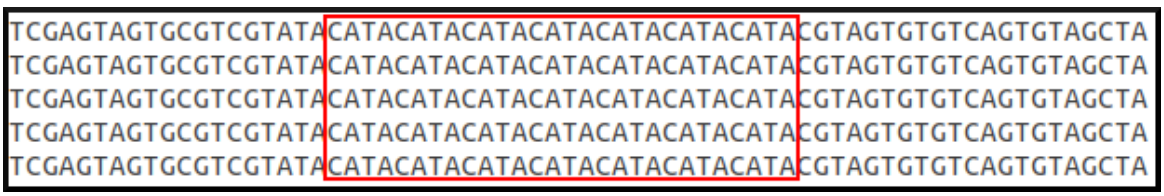
b)


AAGTGCTGTGATGTATATATCGCTTCGTCCCAGAAGTGCTGTCGTATATTTAAAGGGCTCGTAGTAGTAG
AAGTGCTGTGATGTATATATCGCTTCGTCCCAGAAGTGCTGTCGTATATTTAAAGGGCTCGTAGTAGTAG
AAGTGCTGTGATGTATATATCGCTTCGTCCCTGAAGTGCTGTCGTATATTTAAAGGGCTCGTAGTAGTAG
AAGTGCTGTGATGTATATATCGCTTCGTCCCTGAAGTGCTGTCGTATATTTAAAGGGCTCGTAGTAGTAG
ATCGCTTCGTCCCTGAAGTGCTGTCGT

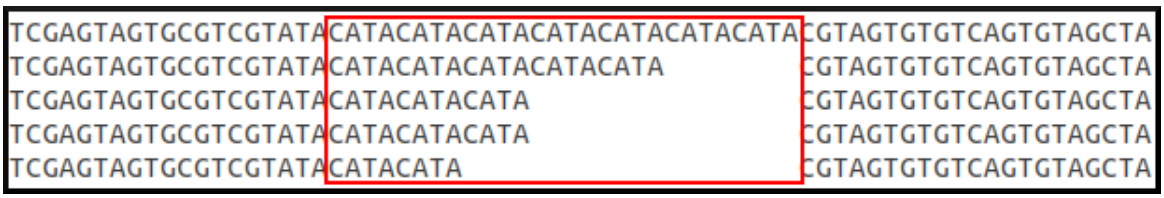
c)


PRIMER
MICROSATELLITE
ACGGCTTAGTCGTTTATATAGCGCGTAGTGACGTGCTGTAGCTAGTTGCTCGCGGAAGGGGGGAAGCG
ACGGCTTAGTCGTTTATATAGCGCGTAGTGACGTGCTGTAGCTAGTTGCTCGCGGAAGGGGGGAAGCG
ACGGCTTAGTCGTTTAT
CGCGGAAGGGGGGAAGCG
ACGGCTTAGTCGTTTATATAGCGCGTAGTGACGTGCTGTAGCTAGTTGCTCGCGGAAGGGGGGAAGCG
ACGGCTTAGTCGTTTATATAGCGCGTAGTGACGTGCTGTAGCTAGTTGCTCGCGGAAGGGGGGAAGCG

d)


TCGAGTAGTGCGTCGTATACATACATACATACATACATACATACATACGTAGTGTGTCAGTGTAGCTA
TCGAGTAGTGCGTCGTATACATACATACATACATACATACATACATACGTAGTGTGTCAGTGTAGCTA
TCGAGTAGTGCGTCGTATACATACATACATACATACATACATACATACGTAGTGTGTCAGTGTAGCTA
TCGAGTAGTGCGTCGTATACATACATACATACATACATACATACATACGTAGTGTGTCAGTGTAGCTA
TCGAGTAGTGCGTCGTATACATACATACATACATACATACATACATACGTAGTGTGTCAGTGTAGCTA

e)


TCGAGTAGTGCGTCGTATACATACATACATACATACATACATACATACGTAGTGTGTCAGTGTAGCTA
TCGAGTAGTGCGTCGTATACATACATACATACATACATA
CGTAGTGTGTCAGTGTAGCTA
TCGAGTAGTGCGTCGTATACATACATACATA
CGTAGTGTGTCAGTGTAGCTA
TCGAGTAGTGCGTCGTATACATACATACATA
CGTAGTGTGTCAGTGTAGCTA
TCGAGTAGTGCGTCGTATACATACATA
CGTAGTGTGTCAGTGTAGCTA

7.4 Appendix 4 - Example costs of microsatellite and SNP analysis for population genetic analysis

Table S1. Reagent and service costs relating to the microsatellite analysis (20 markers) and RAD-Seq / SNP analysis of 96 *H. gammarus* samples.

Reagents & Services for Microsatellite Analysis		**Reagents & Services for SNP Analysis**	
Reagent / Service	**Cost (per 96 samples)**	**Reagent / Service**	**Cost (per 96 samples)**
DNA Extraction: Promega Wizard SV 96 DNA Purification System	£173	DNA Extraction: Promega Wizard SV 96 DNA Purification System	£173
PCR Primers: Custom oligos required for amplification of 20 microsatellite markers	£760	PCR Primers: Nextera XT Index Kit v2 (96 indexes)	£827
PCR Reagents: Qiagen Type-it Microsatellite PCR Kit (inc. extra 20% for optimisations)	£305	NGS Library Preparation Reagents: New England Biolabds NEBNext Ultra II DNA Library Prep Kit for Illumina	£1,811
Capillary Electrophoresis	£1,440	Sonicator Tubes: Covaris microTUBE AFA Fiber Pre-Slit Snap-Cap 6*16mm	£383
Size Standard: Thermo Fisher Scientific GeneScan 500 LIZ dye Size Standard	£172.8	NextSeq Flowcell: Illumina NextSeq 500/550 Mid Output Kit v2 (300 cycles)	£1,721
		Streptavidin Capture Beads: Invitrogen DYNAL MyOne Dynabeads Streptavidin C1	£337
		SbfI Restriction Enzyme: New England Biolabs SbfI-HF	£68
TOTAL	**£2,850.8**	**TOTAL**	**£5,320**